SEQUENCE STABILITY OF THE *APC* GENE: THE ROLE OF DNA REPAIR MECHANISMS IN COLON CARCINOGENESIS

Charlotte Louise Turnbull

Thesis submitted for degree of Doctor in Philosophy

The University of Edinburgh

2007

Dedication

This thesis is dedicated to Paul

Thank-you for everything, I really couldn't have done this without you

Declaration

I declare that this thesis was composed entirely by myself and that the research presented is my own unless otherwise stated.

Charlotte Louise Turnbull

April 2007

The following abstracts derived from the research presented in this thesis are published:

Turnbull, C., Bacon, A., Dunlop, M.G. and Farrington, S.M. (2005) Influence of defective DNA mismatch repair on mutation frequency of *APC* gene. Biochemical Society Transactions, 33(4): P008.

Turnbull, C., Bacon, A., Dunlop, M.G. and Farrington, S.M. (2005) *APC* mutation frequency: the effect of DNA repair. Familial Cancer, 4(1): P145.

Abstract

Mutations are common events in the human genome, arising by a variety of mechanisms (i.e. slippage or mispairing during DNA replication, and via mutagens). This is further complicated by the fact that there may be inherent sequence-specific instability or genomic instability, and defective DNA repair may compound this. *APC* is a key "gatekeeper" gene in colorectal tumourigenesis, with the majority of cases harbouring inactivating mutations. To study the role of instability in colorectal cancer, this project focussed on investigating any inherent sequence-specific mutation frequency of the *APC* gene. I used an approach where clonal selective advantage imparted by *APC* mutation is avoided, and have utilised a number of cancer and non-cancer cell lines, and patient tissue samples. As deficient DNA repair has been implicated in a number of CRC conditions, I have also investigated the effect of defective DNA mismatch repair (MMR) and base excision repair (BER) on *APC* mutation frequency and spectrum. In addition, I have also begun to develop a novel technique for directly measuring BER activity in live cells.

I show elevation of mutation frequency in the *APC* gene compared to control gene sequences. Polymorphism and evolutionary data indicate that *APC* is likely to be subject to a high neutral mutation rate and purifying selection. Furthermore, detailed analysis of *de novo APC* mutations has not implicated the involvement of any alternative mutational mechanisms such as CpG island methylation. Additional work is required to further characterise the possible sequence instability observed in these normal cells.

In patient samples a striking effect of germline *MUTYH* mutation on *APC* mutation spectrum was observed in tumour samples, and the consequence of bi-allelic inactivation of *MUTYH* on somatic inactivating mutations in *APC* and *K-ras* was clear.

The findings presented in this thesis provide new understanding of early molecular events during tumourigenesis, and how they are affected by defects in DNA repair systems, in particular BER and MMR.

Acknowledgements

Firstly, and most importantly, I wish to thank my supervisor, Dr Susan Farrington, for all her support and guidance over the last three years. Thank-you for your patience with all my confusing results, hundreds of questions, practice talks, and thesis drafts – your endurance is truly incredible! Most of all, your ability to give me confidence in the work that I have produced, is something that I hope I will be able to give to the pupils that I will teach in the future.

Thanks also to Prof. Malcolm Dunlop for providing the samples I have needed to complete this work, and for your words of wisdom during lab meetings and progress meetings.

A massive thank-you to Agnes Gallacher and Stephen Brown in technical services, there's no way I could have done this mountain of mini-prepping and sequencing without you – you'll be glad to see the back of me I'm sure!

Last, but by no means least, thanks to all of the 'E4 Bums' past and present, especially those in the big office. Not only have you made every single day of this PhD enjoyable, you've also been brilliant friends throughout.

Abbreviations

A	Adenine
aa	Amino acid
ACF	Aberrant crypt foci
APC	Adenomatous polyposis coli
APRT	Adenine phosphoribosyltransferase
APS	Ammonium persulphate
ATP	Adenosine tri phosphate
BAX	BCL2-associated X protein
BER	Base excision repair
bp	Base pairs (DNA)
BSA	Bovine serum albumin
BRAF	V-RAF murine sarcoma viral oncogene homolog
BUB1 protein	Homologue of the yeast budding uninhibited by benzimidazoles 1
C	Cytosine
cDNA	coding deoxyribonucleic acid
CHRPE	Congenital Hypertrophy of the Retinal Pigment Epithelium
CIMP	CpG Island Methylation Phenotype
CIN	Chromosomal instability
CMV	Cytomegalo Virus
C-MYC	Cellular Myelocytomatosis viral oncogene homologue
COX	Cytochrome oxidase
CRC	Colorectal cancer
CTNNB1	Gene encoding for β-catenin
dH_2O	Distilled water
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleoside tri-phosphate
DSB	Double strand break
EBV	Epstein barr virus

E.coli	*Escherichia coli*
EDTA	Ethylenediaminetetra-acetic acid
FACs	Fluorescent assisted cell sorting
FAP/AFAP	Familial adenomatous polyposis coli/Attenuated
FBS	Fetal bovine serum
FISH	Fluorescence in situ hybridisation
FITC	Fluorescein Isothiocyanate
FGF(R)2	Fibroblast growth factor (receptor) 2
FL1	Fluorescence intensity
G	Guanine
(E)GFP	(Enhanced) Green fluorescent protein
GI	Gastrointestinal
GPD1	Glycerol-3-phosphate dehydrogenase 1
GSK3β	glycogen synthase kinase 3 beta
H_2O_2	Hydrogen peroxide
Hb	Haemoglobin
HEPES	N-2 hyroxyethylpiperazine-N'-2-ethanesulfonic acid
HNPCC	Hereditary non-polyposis colorectal cancer
HPRT	Hypoxanthine guanine phosphoribosyl transferase
IGFR2	Insulin growth factor receptor 2
Ile	Isoleucine
Kb	Kilobase
KRAS	Kirsten rat sarcoma viral oncogene homologue
Lbl	Lymphoblast cell line
LOH	Loss of heterozygosity
LP/SP	Long patch/short patch
MAP	MYH-associated polyposis
MAPK	Mitogen-Activated Protein Kinase
MBD4	Methyl-CpG binding domain protein 4
MCR	Mutation cluster region
MD	Laboratory sample identification
MEF	Mouse embryonic fibroblast

Met	Methionine
MgCl	Magnesium chloride
MIN	Microsatellite instability
MLH1	Homologue of the bacterial MutL mismatch repair protein
MLH3	Homologue of the bacterial MutL mismatch repair protein
MLPA	Multiplex Ligation-dependent Probe Amplification
MMR	Mismatch repair
MNNG	N-methyl-N'-nitro-nitrosanguanidine
MNU	N-methyl-N-nitosourea
mRNA	messenger ribonucleic acid
MSH2	Homologue of bacterial MutS mismatch repair protein
MSH3	Homologue of bacterial MutS mismatch repair protein
MSH6	Homologue of bacterial MutS mismatch repair protein
MSI	Microsatellite instability
MSI-H	High-level microsatellite instability
MSI-L	Low-level microsatellite instability
MSS	Microsatellite stable
MTH	Homologue of bacterial MutT base excision repair protein
MUTYH	Gene encoding for MYH
MYH	Bacterial MutY homologue
NM/B	Normal mucosa/blood sample
NaCl	Sodium chloride
NaOAc	Dodium acetate
NAT	N-acetyltransferase
NSAID	Non-steroidal anti-inflammatory drug
Nt/nt	Nucleotide
NF1	Neurofibromin 1
NF-κB	Nuclear factor-κB
OGG1	8-oxoguanine DNA glycosylase
8-OxoG	8 oxo7,8-dihydro2'deoxyguanosine
PAX6	Paired box gene 6
PBS	Phosphate buffered saline

PCNA	Proliferating cell nuclear antigen
PCR	Polymerase chain reaction
PI	Propidium iodide
PMS1	Homologue of bacterial MutL mismatch repair protein
PMS2	Homologue of bacterial MutL mismatch repair protein
PTT	Protein truncation test
rpm	Revolutions per minute
RNA	Ribonucleic acid
SDS	Sodium dodecyl sulphate
SSC	Side scatter
SSCP	Single strand conformation polymorphism
T	Thymine
TAE	Tris Acetate EDTA buffer
TAF1B	TATA box-binding protein associated factor 1B
TCF-4	T-cell transcription factor 4
TEMED	N, N, N', N' tetramethyl-1-2-diaminomethane
TGFβR2	Transforming growth factor beta type 2 receptor
Tris	Tris(hydroxymethyl)aminomethane
TV	Trypsin versene
UTR	Untranslated region
UV	Ultra violet
VHL	Von Hippel-Lindau tumour suppressor gene
WT	Wild-type
X-gal	5-bromo-4-chloro-3-indolyl-β-D-galactosidase

Table of Contents

Chapter 1 Introduction

Chapter 2 Materials & Methods

Chapter 3

Instability of *APC* Gene sequences and the influence of Defective DNA Repair

Chapter 4

The Influence of Defective DNA Repair on Somatic Inactivating Mutations in *APC* and *KRAS*

Chapter 5

In silico Analysis of Genome Mutation Frequencies

Chapter 6

Measurement of DNA Base Excision Repair in Live Cells

Chapter 7 Discussion and Summary

Chapter 8 Bibliography

Appendix A

Appendix B

Appendix C

List of Tables and Figures

Chapter 1

Introduction

1.1 Overview of colorectal cancer

In developed countries the incidence of colorectal cancer (CRC) is second only to lung cancer (Pisani *et al.*, 1999). Data from the American cancer society (www.cancer.org) indicates that in the USA in 2006 there will be 148,610 new cases of colorectal cancer and 55,170 deaths.

In the UK, the population lifetime risk of developing a large bowel malignancy is 1 in 25 (Dunlop, 1997). Each year CRC causes more than 20,000 deaths in the UK (Farrington and Dunlop, 2004). Locally, in 2003 CRC was the third most frequently diagnosed cancer in both sexes in Scotland, accounting for 14.7% of all male and 11.3% of all female cancers (www.isdscotland.org). CRC is associated with a relatively poor survival rate; in the period 1997-2001 approximately 50% of patients survived five years post-diagnosis (compared with ~77% of breast cancer sufferers) (Scottish Health Statistics, 2006).

Diet and lifestyle appear to have a large influence as to whether an individual develops CRC, with industrialised countries having an incidence that is fifteen times higher than in developing countries (Parkin *et al.*, 1999). The majority of cases of CRC are sporadic, but there is also a heritable component. A malignant tumour arises as a result of multiple interactions between environmental factors and personal genetic factors. Individuals with a first degree affected relative have been shown to have a three fold increased risk of developing adenomatous polyps (pre-malignant lesions) compared with a control population (Bonelli *et al.*, 1988). Epidemiological evidence of a genetic impact on CRC has also been provided by a large Scandinavian twin study (Lichtenstein *et al.*, 2000). A statistically significant effect from genetic

factors on the risk of acquiring CRC was observed, with heritability estimated at 35% (95% CI 10-48%).

CRC is clearly a major public health problem, with the estimated annual cost of treating confirmed cases in Britain exceeding £200m (Dunlop, 1997). Research to elucidate the causes of CRC, and potential preventative therapies are of the utmost importance. There are a number of different areas of research that are thought to have the potential to lead to future developments in the treatment of CRC. These include the identification of new susceptibility/modifier genes, and genetic alterations in tumours, peripheral blood, lymph nodes and bone marrow that will predict prognosis. These will ultimately allow the development of drugs for treatment and prevention, including the potential use of gene therapy to target treatment to tumour cells, and the use of population based genetic screening to target high-risk individuals (Dunlop, 1997).

1.2 Environmental factors in Colorectal Cancer

The high incidence of CRC in western populations is believed to be primarily due to diet and environment. In 1990 the area with the highest estimated rate of colorectal cancer mortality was Australia and New Zealand (30.3 deaths per 100,000 population). Conversely, the area with the lowest estimated rate was middle Africa (0.8 deaths per 100,000 population) (Pisani *et al.*, 1999). Further evidence comes from migrant populations from low incidence countries, who on adoption of the diet of a high incidence country, soon (within one generation) develop the same rate of CRC as natives of that country (Pisani *et al.*, 1999). As such, CRC is a highly preventable form of cancer, and much research has been carried out to deduce what environmental (diet and lifestyle) factors might increase the risk of developing CRC; reviewed in (Gertig and Hunter, 1998) and (Gatof and Ahnen, 2002). Comparisons of diet and CRC incidence amongst different populations report a positive relationship between the high consumption of meat, protein, and fat, and an inverse relationship with the consumption of plant fibre. Table 1.1 provides a summary of the findings from the most significant case-control and cohort studies into the effect of environmental factors on CRC carried out to date.

Table 1.1 Summary of case-control, and cohort studies investigating the relationship between diet/lifestyle factors and the incidence of CRC, adapted from review by (Gatof and Ahnen, 2002).

Variable	Effect	Notes
Obesity	Positive	Effect stronger in males, BMI>30 gives twofold increased risk
Physical Activity	Negative	40-50% reduction in risk - independent of BMI and dietary intake
Carbohydrate - refined	Positive	Diet high in simple sugars may also be deficient in protective nutrients.
Carbohydrate-starch	Negative	Resistance may be due to fermentation in large bowel
Fat	Positive	Relative risk of 1.2 in patient's with highest fat intake
Meat	Positive	Carcinogenic heterocyclic amines may be formed during cooking.
Fibre	Negative	Increases stool bulk and water, reducing gut transit time.
Fruits & Vegetables	Negative	Most consistently seen dietary relationship
Alcohol	Positive	Modest risk
Calcium & Vitamin D	Negative	
Selenium	Negative	May induce apoptosis in a similar way to Non steroidal anti-inflammatory drugs (NSAIDs)
Vitamin E	Negative	
Carotenoids	Negative	
Folic Acid	Negative	May counteract positive effect of alcohol
Vitamin C	Inconclusive	
Tobacco	Positive	2 to 4 fold increase in risk

There has been much evidence to suggest that long-term Non-steroidal anti-inflammatory drug (NSAID) intake can reduce the risk of developing CRC; reviewed in (Williams *et al.*, 1999). Additionally, adenomas have been shown to regress after oral and rectal administration of NSAIDs (Giardiello *et al.*, 1993). The anti-neoplastic effects of NSAIDs, including aspirin, may be due to their inhibition of Cyclooxygenase (COX) enzymes (Williams *et al.*, 1999). Cox-2 has been shown to be over expressed in 80-85% of adenocarcinomas (Eberhart *et al.*, 1994). An alternative mechanism has been proposed that involves the NF-κB pathway. Activation of NF-κB by aspirin is a pro-apoptotic event (Stark *et al.*, 2001), mediated by the sequestering of its p65 component within the nucleolus (Stark and Dunlop, 2005). Unfortunately, the long-term use of aspirin has side effects, including gastrointestinal bleeding. Research is therefore currently looking into alternative NSAIDs such as Sulindac (Matsumoto *et al.*, 2006).

These dietary, lifestyle and chemopreventative factors play a role along side the genetic pathways that lead to sporadic CRC (see section 1.3). The evidence for an environmental basis for CRC is often unclear, with different studies of the same factor showing contradictory affects. Therefore, gene-environment studies that consider both variables may provide a better insight into the role of environmental factors and how they effect different genetic states. For example, dietary factors that have previously been shown to be associated with colorectal cancer risk have been investigated to specifically determine their influence on the occurrence of somatic inactivating mutations in the *APC* gene (Diergaarde *et al.*, 2003). The study found that vegetable consumption was protective against both *APC* mutation-positive and mutation-negative CRC. On the other hand, alcohol consumption was inversely associated with *APC* mutation-positive CRC, whilst meat, fish and fat consumption were positively associated with mutation-positive CRC. Also, the effect of low penetrance polymorphisms in enzymes involved in the metabolism of colon carcinogens has been studied. N-Acetyltransferase (NAT) is involved in the metabolism of cigarette-related carcinogens, and its activity is determined by polymorphisms that cause people to be classified as either 'fast' or 'slow' acetylators. Rapid acetylation at the *NAT* locus has been associated with an increased

risk of developing CRC, due to the activation of substrates that are carcinogenic specifically in the colonic epithelium during rapid, but not slow acetylation (Gertig and Hunter, 1998).

Hereditary colorectal cancer syndromes (including FAP, MAP and HNPCC, see sections 1.5 and 1.6) account for about 5% of all CRCs (Fearnhead *et al.*, 2002). Clearly, this falls significantly short of the estimated level of heritability (~35% as indicated by an epidemiological twin study, (Lichtenstein *et al.*, 2000)). As yet undiscovered genetic susceptibility factors are therefore presumed to account for the remaining 30% of cases caused by heritable factors. The remaining 65% of cases arise as a result of the interplay between environmental factors (some of which have already been identified) and the molecular events occurring within the cell. Lamlum *et al.* estimate that ~10% of individuals with multiple adenomas (but not a classical FAP phenotype) have inherited *APC* variants (i.e. subpolymorphic missense mutations, or mutations at the extreme 5' or 3' end of the gene) that may act alone, or via interaction with other environmental/genetic factors to cause CRC (Lamlum *et al.*, 2000a). Even within individuals with the same CRC syndrome (genotype) great differences can be seen in phenotype. This is due to the vast number of environmental modifying factors involved, and the differences in underlying genetic make-up between individuals (Ilyas *et al.*, 1999). To help understand these relationships further, mouse models can be useful, and these will be discussed in detail in section 1.9.

1.3 Selection and Mutation in Tumourigenesis

Colorectal cancer arises as the cumulative effect of mutations that inactivate tumour suppressor genes and activate oncogenes within the cell, thus allowing it to escape growth and regulatory control mechanisms (inactivation of genes can also occur via epigenetic changes, this is discussed in section 1.8). This step-wise progression of mutations facilitates the histological transition from normal mucosa to adenoma, carcinoma and finally metastasis (Fearnhead *et al.*, 2002).

A number of different models have been proposed to explain this accumulation of mutations during tumourigenesis. The most frequently cited model is of clonal evolution. As mutations occur largely in a random fashion, occasionally a mutation will arise that confers a relative survival advantage to the cell (i.e. allows the cell to evade growth controls or apoptosis). Clonal expansion follows, with a rapid increase in the number of cells harbouring the selective advantage. At this early stage, expansion of the developing clone depends on the rate of cell proliferation out-matching the rate of cell death. Therefore, the sequence of acquired mutations in the developing tumour reflects the specific constraints that are rate limiting at different stages during tumour evolution (i.e. rate of proliferation). Subsequently, when another mutation occurs in one of the clones allowing it to escape additional constraints, further clonal expansion can occur. Later mutations tend to have different functions to those occurring early in the tumourigenic process, with a similar order of mutations in most tumours.

An alternative idea to an ordered progression of mutations during the tumourigenic process is the pre-tumour progression model (Calabrese *et al.*, 2004a). It suggests that the accumulation of mutations may in fact precede tumour progression, occurring in normal appearing colonic crypt niche cells. Mutations are proposed to accumulate within the stem cell population, evolving clonally as multiple stem cells compete for dominance. Eventually, a mutation combination will convert a stem cell into a tumour cell, and visible tumour progression ensues. Using the number of microsatellite mutations present in tumour samples, Calabrese *et al.* calculated the

age at which MMR loss had occurred in a series of patients (Calabrese *et al.*, 2004b). Controversially, they deduced that MMR loss was in fact a late (fifth or sixth) step in the tumourigenic process, as opposed to an early event that leads to increased mutation rate, and therefore subsequent oncogenic mutations.

Most recently a third model has been proposed. Using molecular evolutionary methods to quantify diversity in oesophageal adenocarcinoma, Maley *et al.* provide evidence for a multilineage persistence model (Maley *et al.*, 2006). Showing that lesions with greater clonal diversity are significantly more likely to progress to cancer, the model states that variation is a pre-requisite for tumourigenic progression. In the clonal evolution model, when a single lineage dominates a cell population due to the acquisition of an advantageous mutation, clonal diversity is reduced. Thus, diversity implies a lack of clonal evolution. In the multilineage persistence model however, multiple lineages accumulate mutations (increasing diversity). This increases the chance that one lineage will collect all the mutations required for neoplastic transformation, compared with cells that are undergoing sequential selection.

Within the colon, early mutations are presumed to occur within the stem cell populations of the colonic crypts. It is assumed that mutations occurring in differentiated epithelial cells would be lost due to the rapid turnover of cells within the colon, before there is sufficient time for further mutations to accumulate (Leedham *et al.*, 2005). Work by Komarova *et al.* has disputed this idea (Komarova and Wang, 2004). Using mathematical modelling they suggest a role for differentiating cells in cancer initiation. Although differentiated migrating cells are short-lived, there are a large number of them compared with the number of stem cells. Therefore, it is likely that some will accumulate more than one relevant mutation leading to tumourigenesis.

As discussed in more detail in sections 1.6 and 1.11 the loss of 'genomic stability' can provide a permissive environment for mutations leading to tumourigenesis to occur. For example, loss of mismatch repair or base excision repair dramatically

increases the cellular mutation rate of the genome. Microsatellite instability (MSI), arising as a result of the loss of DNA mismatch repair is observed in ~15% of all colorectal cancers (Grady, 2004). MSI generates mutations in important genes that contain microsatellite repeats, such as *TGFβR2, BAX, IGFR2* and *TAF1B* (Markowitz *et al.*, 1995;Ouyang *et al.*, 1997;Rampino *et al.*, 1997b;Kim *et al.*, 2002;Woerner *et al.*, 2005). If MSI is an early event in tumourigenesis it can therefore alter the type and location of subsequent mutations (Olschwang *et al.*, 1997).

Selection pressure represents a major driving force in tumourigenesis. The genetic pathway that a tumour develops along will be influenced by the selection pressures of the tissue in which it develops. Likewise, different tumours in the same patient are likely to grow at different rates because different oncogenic genotypes do not confer identical selective advantages (Kinzler and Vogelstein, 1996).

There are many stages in the progression of CRC, from normal colonic epithelium to metastasis. The first change is the formation of aberrant crypt foci (ACF's), small areas of epithelium with irregular glandular architecture. Histologically, an adenomatous polyp or adenoma is derived from an ACF. The presence of many hundreds of polyps is characteristic of FAP (see section 1.5), where there are too many to be removed surgically, and so typically the whole colon is removed. The chance of developing polyps also increases with age in non-FAP individuals; by the age of 70 years, approximately half of individuals living in western populations will have developed polyps (Fodde *et al.*, 2001b). If adenomatous polyps are left untreated, cells show increasing levels of dysplasia, followed by a malignant change that results in local invasion and eventually metastasis to distant sites (Vogelstein *et al.*, 1988). Figure 1.1 shows the simplified genetic model of multi-step carcinogenesis within the colon.

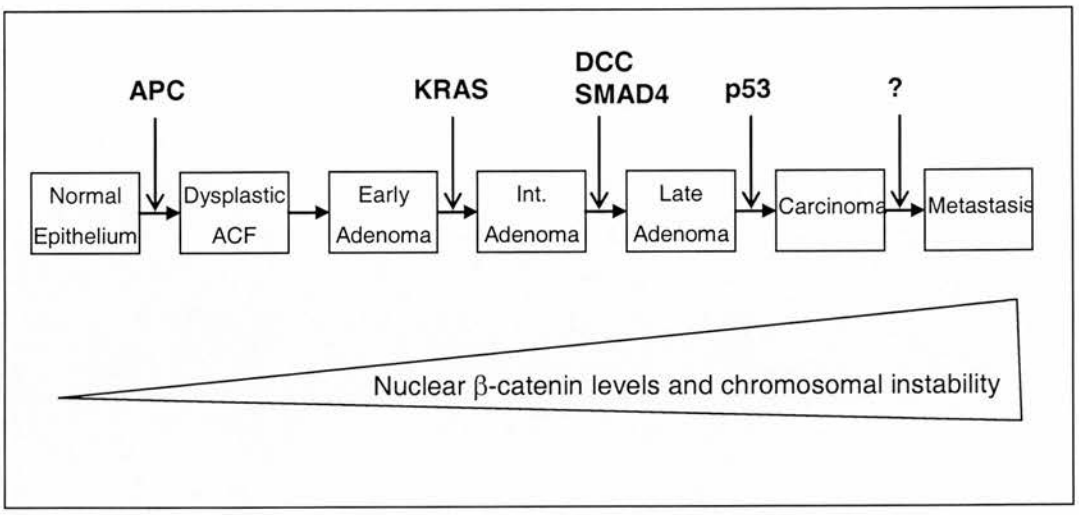

Figure 1.1 Progression of colorectal cancer, with the predicted number of mutational 'hits' required to progress from healthy intestinal tissue to a carcinoma/metastasis. *APC* mutations are thought to initiate the neoplastic process, and are sufficient for colorectal tumours to grow to at least 1cm in diameter (Lamlum et al., 2000b). *KRAS* oncogene mutation is an important early genetic alteration (Vogelstein *et al.*, 1988). Single base pair mutations in this gene can transform cultured cell lines *in vitro* (Land *et al.*, 1983). DNA repair deficiencies will speed up the progression through each stage, by increasing the likelihood of mutations occurring.

1.4 Role of APC

As discussed in section 1.3 inactivation of the *Adenomatous polyposis coli* (*APC*) gene is an early, rate-limiting event in colorectal cancer. This section will look in more detail at the many functions of the APC protein. The *APC* gene was identified as the gene responsible for FAP in 1991 (Groden *et al.*, 1991). It is a large (8535 bp) gene, located at chromosome 5q21, and coding for a 2843 amino acid multifunctional protein. The gene consists of 15 main exons (21 if all alternative exons are included), with exon 15 comprising over 75% of the coding sequence (Bright-Thomas and Hargest, 2003) (Figure 1.2). APC is expressed in a variety of epithelial and mesenchymal cells, including colon, breast, lung and lymphoid tissue (Sieber *et al.*, 2000). There is known tissue specific expression of alternative transcripts, with at least five different forms of 5' non-coding sequences generated by alternative splicing (Horii *et al.*, 1993). Internal translation initiation is also possible at codon 184 in genes mutated at the 5' end of the gene (such individuals often have an attenuated FAP phenotype, see section 1.5) (Heppner *et al.*, 2002).

The APC protein has roles in the control of WNT signalling (Figure 1.3), cell adhesion and migration, apoptosis, and chromosome segregation at mitosis (Fodde *et al.*, 2001b)((Fodde *et al.*, 2001a).

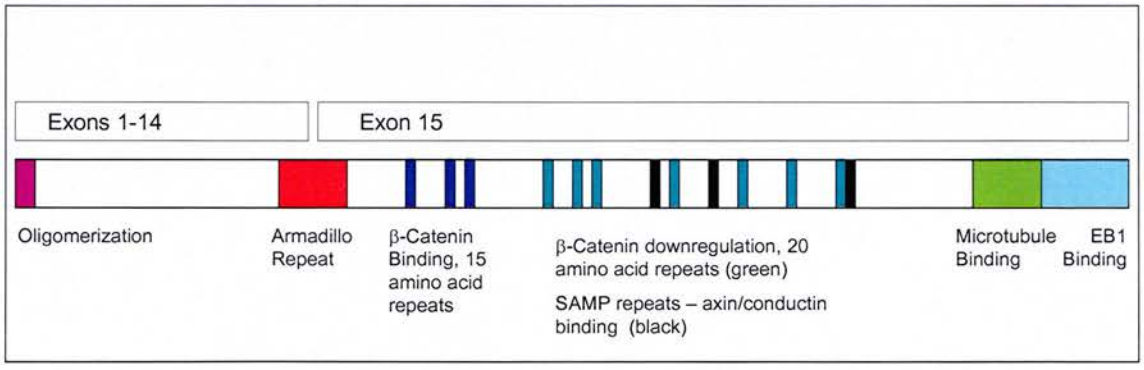

Figure 1.2 Schematic representation of the gene organisation and protein structure of APC.

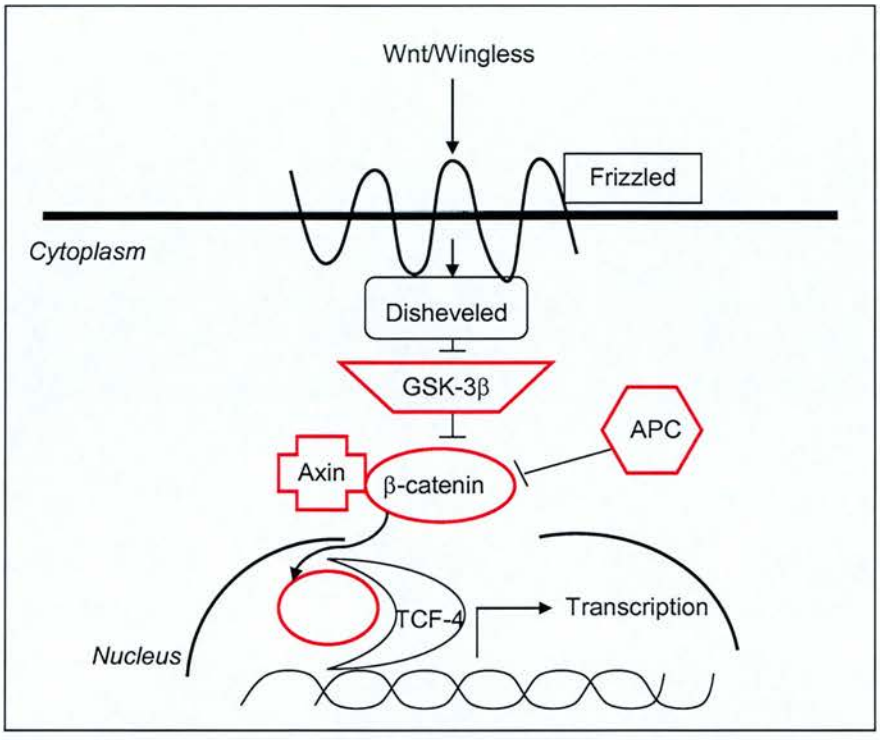

Figure.1.3 The Wnt signalling pathway. When the external signal WNT/Wingless binds to its cell surface receptor frizzled, it causes inactivation of the protein kinase glycogen synthase kinase 3β (GSK3β) (the WNT signal plays a role in intercellular communication, controlling cell fate decisions). GSK3β, Axin, β-catenin and APC form a large protein complex within the cytoplasm (shown in red). In the absence of WNT signalling, GSK3β phosphorylates APC and β-catenin within the complex, signalling the ubiquitin mediated degradation of β-catenin. APC acts as a scaffold protein for the complex, and when it is mutated, or when there is a WNT signal, β-catenin is no longer down-regulated, and is free to translocate to the nucleus and activate TCF-4 (Nathke, 1999). TCF-4 (T-cell factor-4) is a member of the high mobility group (HMG) box transcription factors, activating genes involved in cell proliferation and apoptosis, including those coding for Cyclin D1 and the oncogene *C-myc*. Activating mutations in the oncogene *CTNNB1,* and *AXIN* also cause constitutively activated TCF-4, but these occur very rarely in CRCs compared with *APC* mutations. *APC* and *CTNNB1* mutations are thought to be mutually exclusive (Sparks *et al.*, 1998).

β-catenin also has many roles in the cell, including those involved in tissue architecture and polarity; β-catenin provides the link between E-cadherin, and α-catenin at adheren junctions, that form and maintain the contact between epithelial cells. An increase in the free cytoplasmic pool of β-catenin therefore also increases cell migration.

Evidence suggests that APC has a role in the control of chromosome segregation, a process that ensures that the correct numbers of chromosomes are present in each daughter cell after mitosis. This is of interest because the majority of CRCs are aneuploid (Giaretti *et al.*, 2004;Cardoso *et al.*, 2006). This role of APC in chromosome segregation and stability, may therefore partly explain the aneuploid phenotype in CRC. APC accumulates at the chromosome kinetochore where it facilitates the binding of spindle microtubules (Fodde *et al.*, 2001b). In the absence of a functional C-terminal APC microtubule-binding domain (see Figure 1.2), often lost due to truncating *APC* mutations, an aneuploid/CIN phenotype may occur (Kaplan *et al.*, 2001) (discussed in more detail in 1.11). Loss of microtubule spindle structure due to truncating *APC* mutations can only be rescued by APC protein that contains the missing microtubule binding domain (Dikovskaya *et al.*, 2004). The microtubule binding role of APC means that it also has a number of other important links with the cytoskeleton. Lack of APC microtubule binding affects cell migration, due to loss of interaction of APC and microtubules, and also because of the dominant effects of isolated N-terminal fragments of APC on the F-actin cytoskeleton (Nathke, 2005). Disruption of the microtubule stabilising role has been offered as an explanation for the role of APC in polyp formation in patients with truncating *APC* mutations. Such cells would have compromised migration within the gut, and will consequently be increasingly exposed to toxins for greater periods of time (Nathke, 1999). It is important to note that the cell adhesion and migration functions of APC protein are closely related, and therefore extremely difficult to separate.

The role of APC in proliferation and apoptosis is under some debate, however a number of studies indicate a role for mutant APC in the characteristic tumour traits of increased cell proliferation, and decreased apoptosis. The APC/β-catenin/TCF-4

signalling pathway controls the expression of the anti-apoptotic protein Survivin. Zhang *et al.* demonstrated that wild type (WT) APC down regulates Survivin expression through TCF-4, allowing APC to regulate the size of stem cell populations within intestinal crypts (Zhang *et al.*, 2001). When *APC* is mutated, control is lost, and stem cell populations are free to expand without control. It is interesting to note that as a component of the multi-protein complex Cohesin, Survivin also localises at centromeres. Sister chromatid cohesion until the metaphase-anaphase transition is essential for post-replicative DNA repair by homologous recombination and equal chromosome segregation. Unregulated Survivin may therefore also contribute to genome instability observed during colon tumourigenesis by disrupting the cohesin protein complex (Morrison *et al.*, 2003). The unregulated transcription of other genes such as *MDR-1*, *C-myc*, and *Cyclin D1* in combination, have also been proposed to give rise to proliferating intestinal cell lineages, where normal programmed cell death pathways are suppressed (Yamada *et al.*, 2000).

Over-expression of WT APC in fibroblasts has been shown to slow the progression of the cell cycle (Baeg *et al.*, 1995). Three possible explanations for this observation have been offered. Firstly, it may be due to the binding of the carboxy terminal end of APC to 'Discs Large Gene' (DLG) (Matsumine *et al.*, 1996). DLG is a tumour suppressor protein thought to be involved in cell cycling. Secondly, it has been proposed that a role in cell cycle progression could be mediated by interaction with both the cytoskeleton and DNA (via the β-catenin/TCF-4 pathway) (Friedl *et al.*, 1996). Thirdly, control could be via the β-catenin/TCF-4 pathway alone. TCF-4 drives transcription of Cyclin D1 which is over expressed in many colorectal cancers, and is rate limiting for progression through the cell cycle (Arber *et al.*, 1996).

The many different domains of APC, as demonstrated in Figure 1.2, result in a diverse range of roles of the protein within the cell. As *APC* mutation is the first and rate limiting event in colorectal tumourigenesis it is likely that the different roles of APC within the cell contribute to colorectal cancer progression at many different stages. APC's role in cell migration may be involved in early polyp formation, while

the inability of mutant APC to decrease levels of intracellular β-catenin is likely to contribute to a variety of stages. Whilst all aspects of the multifunctional APC protein are likely to contribute to colorectal tumour formation, and progression, it is APC's tumour-suppressor role in the regulation of intracellular β-catenin levels, that is thought to be its most important function (Fodde, 2003), Figure 1.3.

1.5 FAP

Familial adenomatous polyposis (FAP) (MIM175100) is an autosomal dominant CRC condition with ~100% penetrance. The frequency of the disease has been estimated to be 1 in 13,528 births (Bisgaard *et al.*, 1994). FAP is caused by a germline inactivating mutation in the *APC* gene, and is characterised by the development of hundreds to thousands of polyps within the colon. Due to this high polyp burden, the age at diagnosis of CRC is 20-50 years, much earlier than in sporadic CRC. Additionally, extra-colonic manifestations can include Congenital Hypertrophy of the Retinal Pigment Epithelium (CHRPE), desmoid, adrenocortical and upper gastrointestinal tumours (Miyaki *et al.*, 1995). In addition to the classical form of FAP, an attenuated form exists (AFAP), this is generally characterised by the development of less than 100 adenomatous polyps, and delayed age at cancer onset (Friedl *et al.*, 1996).

An APC database of all published mutations was made available in 1996 (Beroud and Soussi, 1996), with 700 germline and somatic mutations listed. 98% of mutations within the database were either nonsense (30%) or frameshifts (68%) leading to truncation of the APC protein. Germline mutations were scattered throughout the 5' half of the gene, with two hotspots noted at codons 1061 and 1309. A search of the most recent release of the APC database (Thierry Soussi, 2004) reveals that 978/999 (98%) of listed mutations leading to a FAP phenotype are either nonsense or frameshifts that lead to premature termination. Mutations leading to AFAP tend to occur at the extreme 5' end, i.e. the first four exons of the gene, facilitating internal translation initiation, so that the 5' end of the protein remains intact (Heppner *et al.*, 2002). Mutations have also been reported towards the 3' end of the gene, where desmoid disease is also associated with *APC* mutations (Eccles *et al.*, 2001)

Generally the effects of germline missense mutations in APC are not well understood and they are rarely believed to be causative in FAP patients. Germline missense variants of the *APC* gene have however been studied to determine if they may explain why some individuals are predisposed to colorectal adenomas, but don't have

a clear FAP phenotype (Lamlum *et al.*, 2000a). The *APC* variants I1307K (in Jewish populations) and E1317Q (in Caucasian populations) have been found to be significantly associated with multiple colorectal adenomas, and weakly with CRC. I1307K acts as a pre-mutation, by creating a small, unstable segment of DNA, indirectly causing cancer predisposition by facilitating the development of somatic mutations (Laken *et al.*, 1997;Zauber *et al.*, 2003). Although some evidence suggests that E1317Q has some direct pathogenic effect, how it exerts this effect is unclear. The mutant amino acid may also be acting by increasing the selective advantage of other *APC* mutations. Alternatively, the wild-type amino acid may play a role in maintaining the structure of the β-catenin binding and degradation repeats (Laken *et al.*, 1997;Lamlum *et al.*, 2000a). No other germline missense changes in *APC* have been shown to be positively associated with CRC.

As well as truncating mutations, FAP has also been shown to result from genomic rearrangements/constitutive haploinsufficiency of APC (Su *et al.*, 2000;Sieber *et al.*, 2002;Venesio *et al.*, 2003). In 2002 two different genomic deletions that encompassed the whole of *APC* exon 15 were described in two unrelated individuals (Su *et al.*, 2002). Interestingly, one patient had a classical FAP phenotype, whilst the other presented with AFAP. This study underlines the complexities involved in correlating *APC* mutation with FAP phenotype. A Belgian study in 2005 used Multiplex Ligation-dependent Probe Amplification (MLPA) and Fluorescent In-Situ Hybridisation (FISH) to examine patients with classical FAP in whom no truncating *APC* mutation could be found (Michils *et al.*, 2005). They found that 15% of patients with classical FAP have a large deletion of the *APC* gene, accounting for a significant proportion of those that were previously mutation-negative. There is strong evidence therefore that genomic rearrangement is a common type of *APC* mutation, and partly explains why conventional mutation screening of individual exons by PCR cannot always identify a pathogenic mutation. Additionally, studies have looked in more detail at the *APC* inactivating mutations in CRC; Bertario *et al.* examined the relationship between the severity of FAP and the position of the inherited mutation within the *APC* gene (Bertario *et al.*, 2003). They revealed that *APC* mutation sites can be grouped into three sequential subgroups, to identify FAP

patients at high risk of major extra colonic disease. Their results confirmed previous studies suggesting that the common 5bp deletion at position 1309 is associated with a severe phenotype and early CRC development (Caspari *et al.*, 1994).

The dosage effect, which results from having only one functional copy of *APC* increases an individuals likelihood of developing polyps (Yan *et al.*, 2002), but a second, somatic event must occur for progression to CRC (see section 1.7). This second, somatic event may be loss of heterozygosity (LOH), consistent with Knudson's 2-hit model; the remaining wild-type *APC* allele is lost, and so only 1 mutant *APC* allele remains, or a secondary mutation in the remaining *APC* allele is possible. A third possibility is that mitotic non-disjunction can occur, so that daughter cells become hemizygous for the mutant copy of *APC*. Somatic *APC* mutations will be looked at in more detail in section 1.7.

1.6 Cancers associated with DNA repair defects

As well as mutations in oncogenes (i.e. *KRAS, BRAF* and *C-Myc*) and tumour-suppressor genes (i.e. *APC* and *p53*), a third class of genes, involved in DNA repair and maintaining genomic stability can also contribute to tumourigenesis and susceptibility. Germline mutations in *APC* account for <0.1% of CRC's, although 85% of tumours have somatic *APC* mutation. In contrast, cancers associated with loss of DNA repair systems account for ~15% of all cancers, and are therefore much more prevalent than *APC* in the germline (>5%).

1.6.1 Mismatch repair

Hereditary non-polyposis colorectal cancer (HNPCC) (MIM120435) accounts for 5-8% of CRC cases (Lynch and de la Chapelle, 1999). HNPCC is an autosomal dominant CRC condition caused by a deficiency in post-replicative mismatch repair (MMR). Affected individuals carry one mutant copy of a MMR gene, and as with FAP (section 1.5) further somatic events induce early onset cancer (Zhang *et al.*, 2006), most frequently of the colon, but without the characteristic polyps of FAP. HNPCC is associated with an 80% lifetime risk of developing CRC, and 60% risk of developing endometrial cancer (Dunlop *et al.*, 1996;Aarnio *et al.*, 1999). In addition, HNPCC patients are susceptible to other extra-colonic cancers such as endometrial, ovarian, small bowel, sebaceous, central nervous system, ureter and renal pelvis cancer (Watson and Lynch, 1993).

MMR was first described in *Escherichia coli*, where three proteins; MutS, MutL and MutH act together in a complex to identify DNA mismatches, and initiate repair (Kolodner, 1996). Much of what is understood about MMR mechanisms in eukaryotes is based on studies of the prokaryotic system. MutS is an ATPase that drives mismatch recognition. MutL is also an ATPase, that couples mismatch recognition to MutS by downstream processing steps and MutH (a methylation

sensitive endonuclease) is involved in targeting MMR to the newly synthesised strand (Modrich and Lahue, 1996) (Figure 1.4).

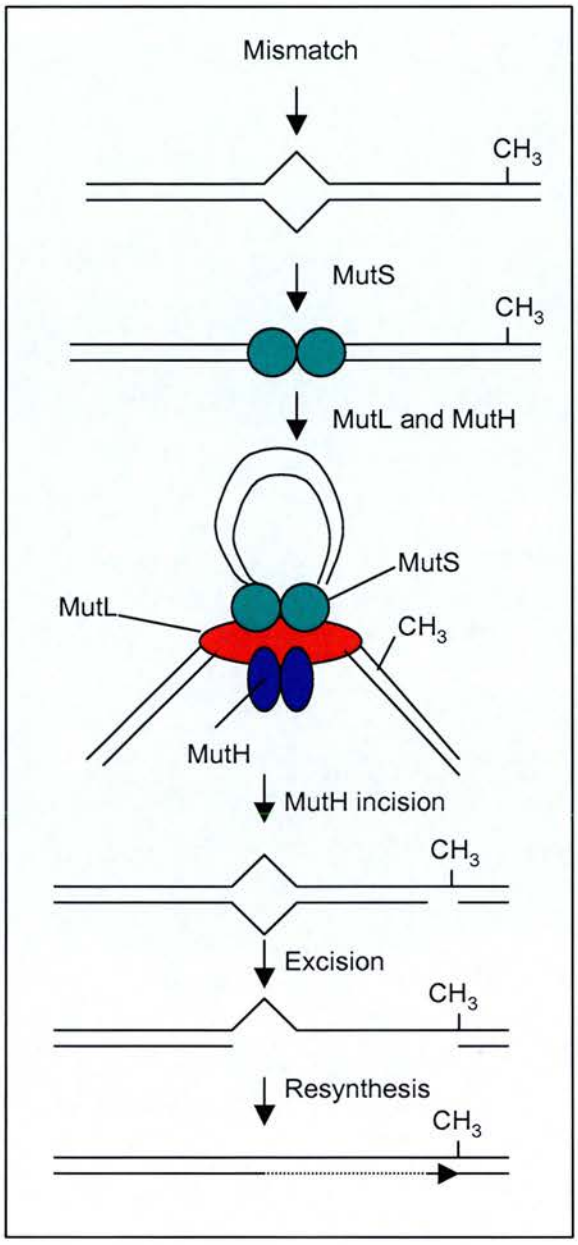

Figure 1.4 Model for mismatch repair of DNA replication errors in *E.coli.* Adapted from (Harfe and Jinks-Robertson, 2000). MutS is a 95kD ATPase that acts as a homodimer to bind base:base mismatches or small insertion:deletion loops (Su and Modrich, 1986). The 25kD endonuclease MutH protein cleaves the unmethylated strand of a hemi-methylated GATC dam methylation site, thereby introducing a nick in the nascent strand for exonucleolytic removal and resynthesis (Modrich, 1991a). The precise role of MutL is less well defined than that of MutS and MutH, however a number of studies have indicated that it serves as a molecular matchmaker, coupling mismatch recognition and downstream MMR events, forming a complex with MutS and enhancing ATP hydrolysis (Modrich and Lahue, 1996;Galio *et al.*, 1999).

The function of eukaryotic MMR is best understood in the budding yeast *S. cerevisiae.* The human MMR system shares genetic and functional similarities with this system. In humans there are now known to be 6 *MutS* homologs (*MSH1-MSH6)* and 4 *MutL* homologs (*MLH1, MLH3, PMS1* and *PMS2),* no MutH homologues have been identified (Harfe and Jinks-Robertson, 2000). Heterodimers form between MSH2 and either MSH3 or MSH6, these are known as MutSβ and MutSα respectively (Acharya *et al.*, 1996). MutSβ is specific for insertion:deletion mismatches whereas MutSα binds preferentially to base substitutions and small insertion:deletion mismatches (Figure 1.5) (Palombo *et al.*, 1996;Umar *et al.*, 1998). The redundancy between MSH3 and MSH6 in the repair of insertion:deletion loops is suggested to explain the prevalence of MSH2 mutations in HNPCC CRCs over mutations in the other two MSH proteins.

In *S. cerivisae* and humans MLH1 is the major MutL homologue and has been demonstrated to form heterodimers with three other MutL proteins (Figure 1.5) (Prolla *et al.*, 1994;Li and Modrich, 1995;Lipkin *et al.*, 2000). Interactions between the MLH1-PMS2 complex and MSH2 complexes have been demonstrated, suggesting that MLH1-PMS2 acts as a molecular matchmaker, co-ordinating activities between MSH complexes and downstream MMR events in role similar to that of MLH1 in bacteria (Figure 1.5) (Prolla *et al.*, 1994;Kolodner and Marsischky, 1999). MLH1-PMS1 heterodimers are referred to as MutLβ, whose role in MMR has not yet been demonstrated. MLH3 interacts with MLH1 (Lipkin *et al.*, 2000) and its homology to the yeast MLH3 protein suggests that it is involved in the repair of a subset of insertion:deletion mispairs recognised by the MSH2-MSH3 complex (Flores-Rozas and Kolodner, 1998). The role of the human MLH1-MLH3 complex remains to be elucidated.

The assembly of MLH and MSH complexes subsequently recruits a number of other proteins that are not specific to MMR activity but are involved in completing the later stages of MMR. These include DNA polymerase δ, replication protein A, proliferating cell nuclear antigen (PCNA), replication factor C (RFC), exonuclease 1,

Fen 1 and the DNA polymerase δ and ε associated exonucleases (Kolodner and Marsischky, 1999).

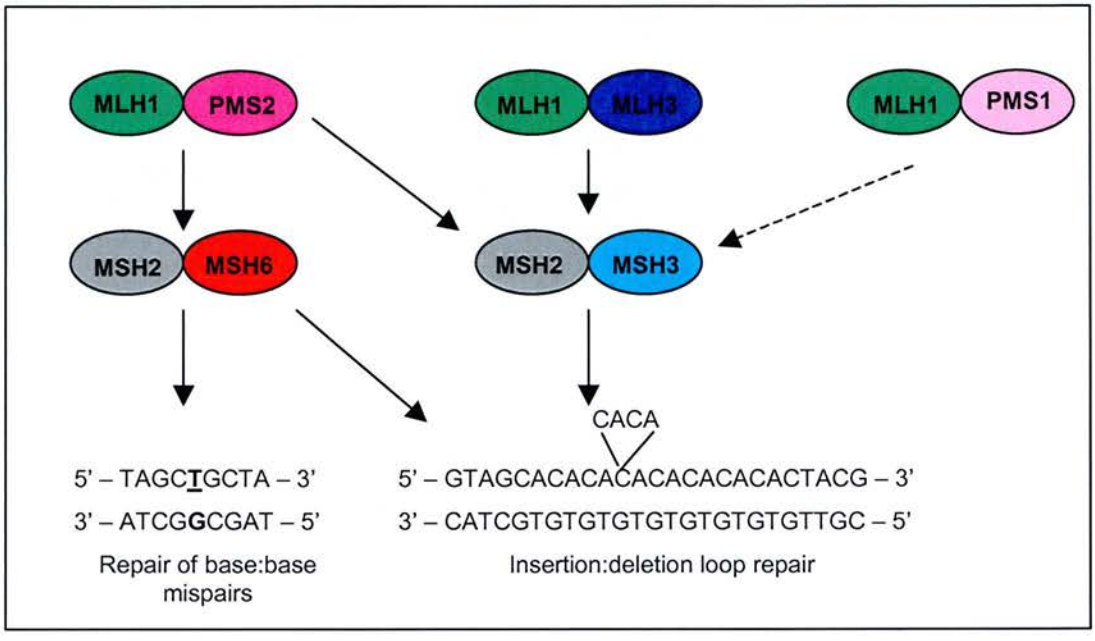

Figure 1.5 Interaction of human MMR proteins in the repair of base:base mispairs and insertion:deletion loops. Adapted from references (Kolodner and Marsischky, 1999;Harfe and Jinks-Robertson, 2000).

The genetic basis for HNPCC was initially reported by Peltomaki *et al.* in 1993, who mapped the first locus to 2p15-16 (Peltomaki *et al.*, 1993). MMR deficiency in sporadic and heritable cancer is usually caused by inactivating mutations (or promoter hypermethylation) mainly in the *MLH1* (~50%), *MSH2* (~40%), or *MSH6* (~10%) genes (Peltomaki, 2001), as these are directly involved in post-replication repair. *MSH3* plays a role in sporadic, whilst *PMS2* and *PMS1* play a minor role in both sporadic and hereditary CRC. The other proteins have different roles i.e. *MSH4* and *MSH5* function in meiosis.

MMR is important in maintaining the fidelity of replication, as well as targeting repair of mispaired bases from replication errors, homologous replication and DNA damage. MMR deficiency increases the rate of mutations in tumour cells by 100 to 1000 times over normal cells (Modrich, 1991b). This is highlighted by the observation of a high frequency of mutations in non-coding microsatellites where there is no apparent selection pressure (Hienonen *et al.*, 2005). A deficient MMR system leads to the accumulation of mutations (base-base mismatches, and insertion, deletion loops) and instability of short tandem repeats (microsatellite instability (MSI)) within the genome. MSI is a consequence of DNA polymerase slippage and consequently strand misalignment at repetitive sequences during DNA replication. If uncorrected, these errors become fixed as a mutation. If post-replicative errors occur within tumour suppressors or oncogenes they significantly increase an individual's risk of developing cancer. In addition MMR deficient cells are hyper-recombinogenic. This is because the MMR system usually imposes a barrier to recombination, enabling the repair of DNA damage caused by agents such as UV radiation, reactive oxygen species, and alkylating agents (Schofield and Hsieh, 2003).

MSI is used as an indicator of MMR deficiency in CRC, with >90% HNPCC cases, and ~15% of sporadic CRC's displaying MSI (Liu *et al.*, 1995;Fearnhead *et al.*, 2002). Inactivating mutations in MMR genes are identified in less than 30% of sporadic CRC's displaying MSI, in the rest, a role of epigenetic silencing of MLH1 is regarded as more important (Oda *et al.*, 2005). Sporadic tumours with a very high

frequency of methylation at the CpG islands in the MLH1 promoter, and at other sites are known as CIMP (CpG Island Methylator Phenotype) positive (Issa, 2004a). In HNPCC and sporadic disease repetitive sequences are found within the important growth control genes *TGFβR2* (Transforming growth factor β-receptor 2) and *IGFR2* (Insulin growth factor-receptor 2), and genes associated with apoptosis i.e. *BAX* (Bcl-2-associated X protein gene) and *Caspase5*. These are useful indicators of MSI (Oda *et al.*, 2005). Some genes show a high frequency of frameshift mutations even at the early adenoma stage of MSI tumourigenesis; 72% of MSI adenomas and 90% of MSI carcinomas show mutations at a *TGFβR2* poly(A) tract (Markowitz *et al.*, 1995;Woerner *et al.*, 2005). *BAX* is a pro-apoptotic gene containing a poly$(G)_8$ tract in exon 3 (Ouyang *et al.*, 1998). The balance between BAX and Bcl-2 (the antagonist of BAX) regulates the apoptotic response of the cell to apoptotic stimuli (Pawlowski and Kraft, 2000). Around 50% of CRCs with MSI have frameshift mutations at the poly$G_{(8)}$ tract of BAX, creating a diminished ability to trigger apoptosis upon receiving a death signal (Rampino *et al.*, 1997a). The loss of these apoptotic response genes in MMR deficient tumours is further compounded by the fact that the MMR proteins also have their own role in the initiation of apoptosis in response to the detection to DNA damage (Zhang *et al.*, 1999). Additionally, MMR deficiency influences subsequent gene mutations during the progression of tumourigenesis. *TCF-4* (see section 1.4) has been shown to be a target gene for frameshift mutations in MMR deficient colon cancers (mostly sporadic cases), due to its sequence structure and role in oncogenesis (Peltomaki, 2001). The BRAF V599E missense change is also associated with MSI colorectal carcinomas (Rajagopalan *et al.*, 2002). BRAF is a downstream actor of KRAS in the MAPK pathway, and plays a key role in the oncogenesis of melanoma and CRC (Davies *et al.*, 2002). Despite the fact that the V599E BRAF alteration occurs only in CIMP tumours with MSI, it does not arise as a direct consequence of deficient MMR. This indicates that MSI caused by epigenetic MLH1 silencing creates a distinct mutational background from that caused by genetic loss of MMR (Lubomierski *et al.*, 2005).

A better prognosis is associated with MMR deficient colorectal cancer, compared with cancers that retain MMR activity (Jass, 2000), although this may be associated

with sporadic MMR deficiency only, as two reports from our own lab indicate that germline MMR gene carriers do not have a survival advantage (Farrington *et al.*, 2002;Barnetson *et al.*, 2006). A micro array study of 12,625 probes revealed that MMR inactivation produces distinct changes in the cellular mRNA pool of tumours and colorectal cancer cell lines, when compared with tumours and cell lines with intact MMR (di Pietro *et al.*, 2005). MMR deficient cancers had reduced transcript levels of several genes that favour tumour progression and increased levels of factors believed to inhibit tumour growth, progression and invasiveness (i.e. Caspase2 and SMAD4). The authors propose that reduced levels of transcription may be due to microsatellite sequences in the 5'UTR (affecting transcription efficiency) or coding region (leading to nonsense mediated decay). In contrast, shortened microsatellite sequences in the 3'UTR of genes that showed increased levels of transcription in MMR deficient cancers were observed, suggesting a possible effect on RNA stability, processing and translation efficiency.

The type of *APC* mutations observed in CRC tumours can differ, depending on whether or not there are defective DNA repair pathways present. There is some evidence to suggest that the *APC* mutation spectrum is affected by the loss of MMR (Huang *et al.*, 1996), with deficient tumours having a higher rate of *APC* frameshift mutations than MMR proficient tumours, and a tendency towards them occurring within mononucleotide repeats. A coupled *in vitro* synthesised protein test was used to look at the truncating mutations occurring at *APC* in patients with MSI tumours (Huang *et al.*, 2004). They found that 58% of cases had *APC* mutations; 69% of which were frameshifts (mostly at mononucleotide repeats) and 31% nonsense (missense mutations were not investigated). They concluded that most mutations arose directly as a result of deficient MMR, and therefore that defective MMR was occurring earlier than *APC* mutation in these patients. Previous reports have however produced conflicting results (Lamlum *et al.*, 2000b), and are discussed further in section 1.7.

BER deficient tumours also show a specificity for *APC* mutation with a propensity for G to T transversions (Al Tassan *et al.*, 2002;Jones *et al.*, 2002), as described in further detail below.

1.6.2 Base excision repair

Heritable mutations in the genes involved with the BER system have recently been implicated in CRC. BER protects cells against the deleterious effects of DNA oxidation, specifically the removal of 8-oxo7,8-dihydro2'deoxyguanine (8-oxoG/8-hydroxyguanine), the most stable and mutagenic product of oxidative DNA damage (Hayashi *et al.*, 2002) (Figure 1.6)

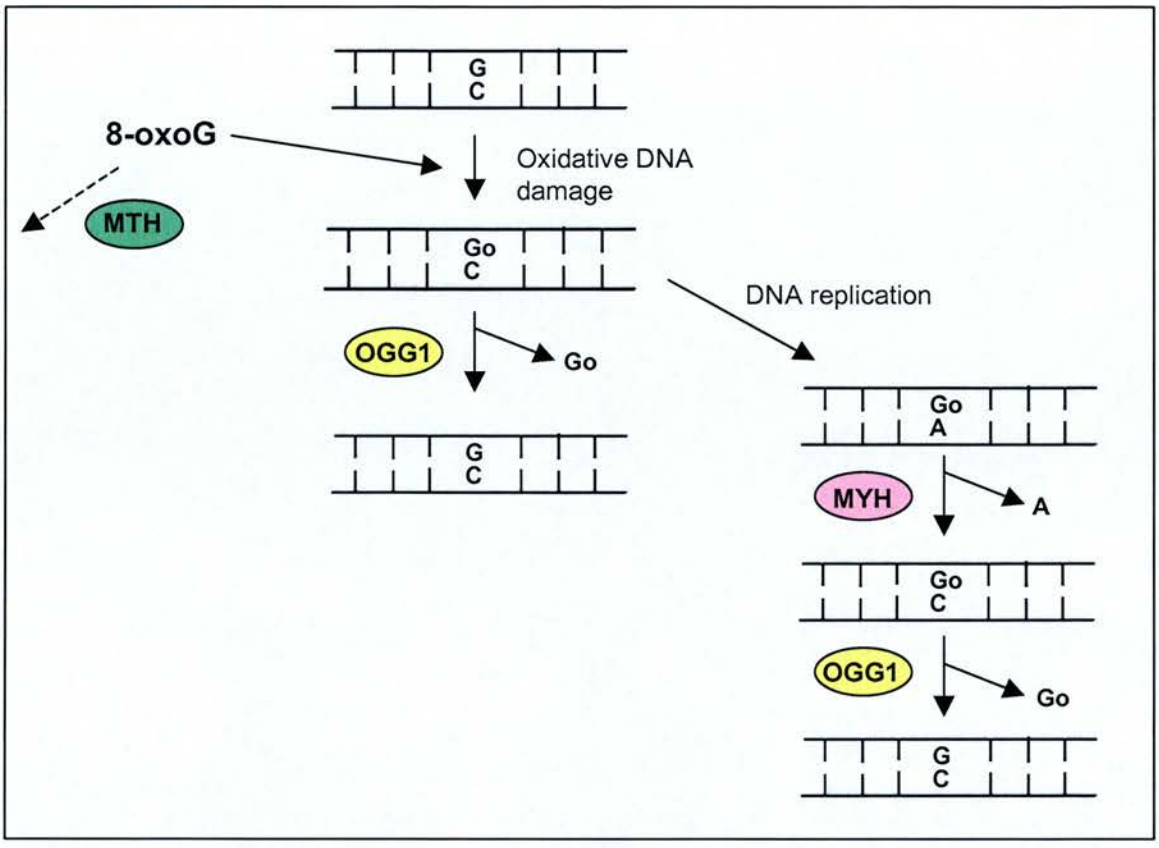

Figure 1.6 A model of base-excision repair in mammalian cells, adapted from (Slupphaug *et al.*, 2003). MTH (*E. coli* mutT homolog) is the first line of defence, acting by removing 8-oxoG from the nucleotide pool via hydrolysation (Fortini *et al.*, 2003). If 8-oxoG escapes this sanitation of the nucleotide pool, it may be incorporated into DNA by polymerases. During DNA replication 8-oxoG can become incorporated opposite C residues in place of G. 8-oxoG may then pair with A at the next round of replication, leading to a G:C to T:A transversion mutation. OGG1 (a specific DNA glycosylase) can remove 8-oxoG from the template strand prior to DNA replication. In cases where 8-oxoG:C lesions are not repaired before the next round of replication, the lesion will become 8-oxoG:A. In these cases the MYH glycosylase protein (the *E. coli* MutY homolog) will then excise the A residue that has become mis-incorporated opposite 8-oxoG during replication, allowing the 8-oxoG:C lesion to be repaired by OGG1. The importance of these repair systems in humans is supported by the finding that inherited variants of MYH at conserved amino acid residues are associated with increased somatic G:C to T:A transversions in colorectal tumours (Al Tassan *et al.*, 2002). Oxidative DNA damage is particularly prevalent in the gut, which is why mutations in the BER genes are thought to be specifically important in colorectal tumours.

Individuals with inactivating mutations in the BER gene *MUTYH,* coding for MYH, have a higher rate of G:C to T:A transversions within the *APC* gene than those with intact *MUTYH* (Al Tassan *et al.*, 2002). Inherited biallelic mutations in *MUTYH*, cause an autosomal recessive trait ('*MYH*-associated polyposis', MAP) that predisposes individuals to the formation of colonic polyps (Sampson *et al.*, 2003) due to an increased mutation rate, and specific mutations in *APC*, *KRAS* (Lipton *et al.*, 2003;Jones *et al.*, 2004) and potentially other, as yet unidentified genes involved in the pathogenesis of colorectal tumours (Halford *et al.*, 2003). Individuals with MAP typically have an attenuated or atypical adenomatous polyposis phenotype (Aretz *et al.*, 2006b). Age at diagnosis is later (~50 years), and patients present with less polyps and without the severe extra-intestinal manifestations associated with typical polyposis (FAP) (Aretz *et al.*, 2006a). There may also be a small increase in risk of developing colorectal cancer for individuals with heterozygous *MYH* mutations (Farrington *et al.*, 2005;Jenkins *et al.*, 2006;Tenesa *et al.*, 2006).

MYH enhances the fidelity of DNA replication and genetic recombination via physical association with replicative proteins, and mismatch repair proteins (Gu *et al.*, 2002). While MMR mutations are seen in ~10-15% of sporadic CRCs, mutations in *MUTYH* do not appear to have a significant affect on sporadic disease. However, allelic loss of 1p loci has been described as an early event in colorectal tumourigenesis (Lothe *et al.*, 1995), and LOH of the remaining allele in *MUTYH* heterozygotes may provide an explanation for the increased risk of developing CRC in these patients (Croitoru *et al.*, 2004b). It has been suggested that the lack of somatic *MUTYH* mutations may indicate that MMR and oxidative repair genes predispose to bowel tumours through fundamentally different mechanisms (Halford *et al.*, 2003). Nonetheless, there is interaction between the MMR and BER pathways. For example, oxidative DNA damage can contribute to frameshift mutations, due to 8-oxoG becoming incorporated opposite C and A residues in repetitive DNA sequences (Macpherson *et al.*, 2005). In turn, MMR reduces the burden of DNA 8-oxoG, since MutSα (a heterodimer of Msh2 and Msh6) is able to recognise 8-oxoG mispairs, and initiate MMR (Larson *et al.*, 2003). Levels of steady-state and H_2O_2-induced DNA 8-oxoG were shown to be increased in *msh2*$^{-/-}$ mouse embryonic

fibroblasts (Colussi *et al.*, 2002;Gu *et al.*, 2002). Thus, deficient MMR contributes to the defective processing of oxidative DNA damage. Cancers showing both MMR and BER deficiency have not been described (although patients carrying both a MMR gene mutation and a heterozygote *MUTYH* mutation have been demonstrated (Niessen *et al.*, 2006)), this is possibly due to the fact that a double deficiency is simply not required for genomic instability, or that cells could not cope with two defective pathways (Lipton *et al.*, 2003).

The protein products of many of the genes discussed above, have been linked to DNA damage-induced cell-cycle arrest, and apoptosis (Heinen *et al.*, 2002). This indicates that as well as causing the primary genomic instability pathway to tumourigenesis, DNA repair mutations also give a second selective advantage to the cell, as was also seen for *APC* mutations (section 1.4). This goes some way to explain why, in mouse models, the MMR deficient $Msh6^{-/-}$ embryonic stem cells, but not $Msh3^{-/-}$ cells display resistance to the mutagen N-Methyl-N'-Nitro-N-Nitrosoguanidine (MNNG). While the heterodimer MSH2-MSH6 is involved in DNA damage recognition, MSH2-MSH3 is not (de Wind *et al.*, 1999), so the Msh6 knockout is unable to detect mutagen induced damage, but the Msh3 knockout retains DNA damage recognition.

The elevated mutation rates observed in cells with deficient DNA repair systems is often referred to as a mutator phenotype, and has been recognised as an early genetic event in sporadic and hereditary cancers (Loeb, 2001). A mutator phenotype (discussed in more detail in section 1.11), combined with selection pressure (as discussed in section 1.3) is believed to facilitate the accumulation of mutations in important regulatory genes, that ultimately leads to the initiation and progression of tumourigenesis.

1.7 Somatic and sporadic *APC* mutations

As discussed in section 1.5, in FAP patients, tumours are initiated by additional somatic mutations affecting the wild-type *APC* allele. The majority (>85%) of somatic *APC* mutations are known to occur within the mutation cluster region (MCR), a region of exon 15 from codons 1286 to 1513, containing the β-catenin binding amino acid repeats (Figure 1.5). In contrast, only 29% of germline mutations are found within the MCR (Miyaki *et al.*, 1995). β-catenin levels have been proposed to be important in deciding where in the gene a second *APC* mutation occurs. Retention of one of the β-catenin down-regulating 20-amino-acid repeats in either of the two *APC* alleles appears to be advantageous during tumourigenesis. This suggests that there is some optimal level of β-catenin signalling that is required for tumour formation, and is termed the 'just-right' signalling model (Albuquerque *et al.*, 2002).

Within the context of cancer biology somatic cell genetics generally refers to the molecular events that occur in cells before and during neoplastic transformation. It is thought that increased knowledge of somatic cell genetics will greatly affect the prognosis of sporadic cancer. To identify common alterations involved in tumour pathogenesis and progression Blaker *et al.* looked for second somatic mutations, and the corresponding immunohistologic phenotype of β-catenin, that followed a known germline mutation in ten different adenomatous polyps from a single FAP patient (Blaker *et al.*, 2003). All polyps had a second somatic mutation; eight frameshifts (three of which were identical), one nonsense mutation, and one missense mutation, six also had chromosomal imbalances. All polyps had enhanced membranous and cytoplasmic β-catenin staining, although only 1 showed nuclear translocation of β-catenin. It is widely accepted that biallelic *APC* inactivation is an early event in tumourigenesis that is necessary, and sufficient for adenoma development (Lamlum *et al.*, 2000b;Fodde *et al.*, 2001b). However, whilst loss of APC function may be a prerequisite, it is not the sole cause of nuclear translocation of β-catenin. Likewise, chromosomal alterations are not required for adenoma development, but are often a consequence of it (Fodde *et al.*, 2001a).

As mentioned previously, ~30-35% of all colorectal cancers are thought to have a hereditary component (Lichtenstein *et al.*, 2000;Fearnhead *et al.*, 2002). The remaining cases are likely to arise as a result of purely sporadic mutations. These occur in the somatic cells of the affected individual, or can occur in the germline of a parent and are subsequently inherited as a *de novo* mutation, as in the case of some sporadic and also potentially *de novo* FAP cases (Farrington and Dunlop, 1999). The majority of sporadic colorectal cancer cases also have inactivating mutations on both *APC* alleles (Nagase and Nakamura, 1993).

Due to a lack of detectable parental origin bias, or parental age effect Ripa *et al.* propose that *de novo APC* germline mutations arise during meiotic division, compared with somatic mutations (and mutations in mosaic individuals) that arise during mitosis (Ripa *et al.*, 2002). However, there is conflicting evidence of a parental origin for *de novo APC* mutations (Ripa *et al.*, 2002;Aretz *et al.*, 2004), with small sample sizes hindering significant conclusions.

The difference in distribution between somatic and germline mutations within the *APC* gene alongside other evidence may suggest that *APC* is susceptible to genetic changes during meiosis, compared with most other genes where changes occurring at meiosis are not observed. A specific role for polymerase β in repair synthesis during the early stages of meiosis rather than mitosis has been proposed (Plug *et al.*, 1997). Additionally, the replication error spectrum of polymerase β has also been shown to contain a significant subset of the *APC* hotspots seen in colorectal cancer (Muniappan and Thilly, 2002). These include a 5bp deletion at codon 1309 (delAAAGA), found to be over-represented in patients with *de novo* mutations (45%) compared with familial mutations (5%) (Aretz *et al.*, 2004). This provides good evidence for a mechanism involving polymerase β errors arising during meiosis to explain how *de novo* germline mutations arise in *APC*. With the possibility that the sequence context of the *APC* gene is increasing its susceptibility to errors arising during polymerase β repair synthesis. An alternative argument could also be that a wide spectrum of germline mutations are occurring, but that are permissive for only a

limited spectrum of somatic second mutations, thus leading to the differences observed in the distributions of somatic and germline mutations.

Further analysis of somatic mutations in the *APC* gene has been performed by Luchtenborg *et al.*, who examined the MCR of *APC* in a Dutch cohort of sporadic colorectal carcinomas (Luchtenborg *et al.*, 2004). They observed that the majority of mutations (61%) were missense point mutations. The relevance of missense mutations in *APC* are not well understood, and more investigation is required to determine which, if any, have a pathogenic effect. Truncating *APC* mutations were seen in 37% of cases – 52% of which also had a missense mutation. Multiple mutations (between 2 and 8) were seen in 55% of cases, and underline the existence of tumour heterogeneity in sporadic CRC. Heterogeneity in sporadic cancers is expected to arise as the result of sub clones harbouring different mutations (Maley *et al.*, 2006), but may also indicate a defect in DNA repair. The presence of a truncating and missense mutation in the tumour DNA of a single patient, if they are biallelic, may indicate pathogenicity of missense mutations, but whether mutations were present on the same allele was not investigated in this study.

1.8 Epigenetic changes in CRC

Epigenetics is an alternative mechanism in carcinogenesis, with most colorectal cancers having some epigenetic abnormalities (Kondo and Issa, 2004). As mentioned previously (section 1.6.1), promoter CpG island methylation, is the most common form of epigenetic change, and results in irreversible inhibition of gene expression. A subset of colorectal tumours have an exceptionally high frequency of methylation of some CpG islands, and are referred to as CIMP positive/high (Issa, 2004b). CIMP positive tumours are associated with a variety of clinical, histopathological and epidemiological characteristics (Samowitz *et al.*, 2005), most notably MSI.

MMR deficiency in sporadic colorectal cancer is known to be caused largely by epigenetic mechanisms. A number of studies have shown that the most common mechanism causing MSI in sporadic colorectal cancer is transcriptional silencing of the *MLH1* gene by methylation of the *MLH1* gene promoter (Lei *et al.*, 2004). Veigl *et al.* also showed that 5 out of 6 MSI cell lines with no detectable MLH1 protein, but wild type coding sequences had promoter hypermethylation of the gene promoter from both parental alleles (Veigl *et al.*, 1998).

As well as the better-documented MLH1 case, the *APC* gene is also known to be silenced by epigenetic mechanisms. Work by Esteller *et al.* in 2000 highlighted the fact that in some non-familial tumours with *APC* mutations there is no second identifiable event, and up to 20% of tumours have no *APC* mutation (Esteller *et al.*, 2000). They showed that the *APC* promoter 1A is hypermethylated, and unexpressed in 18% of sporadic colorectal carcinomas and adenomas. Hypermethylation appears to occur early in colon tumourigenesis, and usually affects intact copies of *APC*.

The *APC* gene contains two promoters, 1A and 1B. The main *APC* transcript is initiated from 1A, and a minor transcript from 1B. Hypermethylation of promoter 1A has been demonstrated in melanoma cell lines, that did not show increased WNT signalling (Worm *et al.*, 2004). It is suggested that this may be due to residual activity from promoter 1B, causing a promotion of cell proliferation, without

compromising invasive capacity. Indeed, promoter 1A hypermethylation is a normal event in the stomach (but not the colon), here virtually all *APC* transcripts result from promoter 1B (Tsuchiya *et al.*, 2000).

1.9 Mouse models

Mouse models are used in the study of CRC to further dissect complicated genotype-phenotype relationships, to look at the interplay between different genotypes, and as preclinical models for the testing of preventative, and therapeutic agents (Fodde and Smits, 2001). Whilst useful, it is important to be aware of the limitations of mouse models. Differences in phenotype and disease progression are often seen between humans and mice with similar genotypes, and so any conclusions drawn from these models must account for these differences.

The multiple intestinal neoplasia (*Min)* mouse has a nonsense mutation at codon 850 of *Apc*, resulting in truncation of the protein product (Moser *et al.*, 1990;Su *et al.*, 1992). Heterozygous Apc^{+}/Apc^{Min} mice have a severe phenotype, developing over 100 intestinal tumours (but not polyps), and rarely surviving past 140 days. Tumours tend to appear in the small intestine, rather than the colon, and due to the severity of disease, and lack of polyps, early events in tumourigenesis are difficult to study in the Min mouse. In the majority of cases, loss of the WT *Apc* allele is the result of the entire loss of mouse chromosome 18 (Luongo *et al.*, 1994), syntenic to human chromosome 5, but ultimately very different to the situation observed in humans.

The Min mouse has been useful in a large number of *in vivo* studies into chemopreventative and therapeutic strategies for colorectal cancer, where the suppression of intestinal adenoma development can be monitored and compared under a variety of conditions. The ability of PEG-3500 (a polyethene formulation) to stimulate apoptosis as a potential chemopreventative treatment was examined by Roy *et al.* They found that clinically relevant levels of PEG-3500 in drinking water could reduce tumour burden and resulted in a 2-3 fold increase in colonic mucosal apoptosis (Roy *et al.*, 2004). Likewise, Sale *et al.* detected a 24% decrease in adenoma load following the administration of the synthetic resveratol analog DMU-212 to Apc^{Min} mice (Sale *et al.*, 2005).

The Min mouse was used to identify a modifier locus, 'modifier-of-Min' (*Mom-1*) (Dietrich *et al.*, 1993). *Mom-1* encodes a secretory phospholipase A2, a gene involved in lipid metabolism and prostaglandin production (MacPhee *et al.*, 1995). In mice it reduces the number of tumours seen in the Apc^{+}/Apc^{Min} mice, and lies in a region of the mouse genome syntenic to human chromosome 1p35-36. This region is known to be lost in many human cancers, including colorectal cancer. However, no correlation has been established for phospholipase A2 gene mutation and the FAP phenotype therefore, there is no evidence in humans that this region acts as a modifier (Farrington and Dunlop, 2004).

The Apc protein in the Apc^{1638T} mouse is truncated at codon 1638, and so retains all β-catenin binding repeats, 3 of the 7 β-catenin down regulating repeats, and one axin/conductin binding domain, but lacks all the C-terminal interaction domains (Smits *et al.*, 1999). Heterozygous, and indeed homozygous Apc^{1638T} mice, show no increase in tumour susceptibility, indicating that the C-terminal end of Apc is not required for tumour suppression (in mice). However, homozygous embryonic stem (ES) cells do show a high degree of chromosomal instability. It is proposed that this indicates that Apc is directly involved in chromosomal stability, but that while chromosomal instability is necessary, it is not sufficient to trigger tumour formation (Fodde *et al.*, 2001a).

The Apc^{1638N} mouse has a 'leaky' allele, producing 1-2% of the truncated protein product. *Apc*+/1638N mice, like Min mice have a distribution of tumours that is different to that seen in humans, with the majority being in the upper GI tract, between the stomach and small intestine (Fodde and Smits, 2001). On a B6 background, 1638N mice are particularly prone to developing extra-intestinal lesions, similar to FAP patients. These include desmoid tumours, cutaneous cysts and a CHRPE-like retinal abnormality (van der, 98 A.D.). The two mouse models, Apc^{1638T} and Apc^{1638N} provide strong evidence that there is haploinsufficiency of the mutant Apc protein, so the leaky allele at 1-2% of normal expression creates a detectable phenotype that is not seen when the same mutant allele is expressed at 'normal', higher levels.

The importance of somatic loss of APC in sporadic CRC has been underlined in the *Apc*580S mouse (Shibata *et al.*, 1997). In this model the LoxP/Cre system is used to somatically inactivate both *Apc* alleles (by deleting exon 14, creating a frameshift), inducing tumour formation within the colorectum extremely rapidly, within 4 weeks of conditional targeting of *Apc*.

The synergistic interaction between the *Apc* gene and defective MMR has also been explored in mice; Msh2 and Mlh1 deficiencies have been shown to cause somatic *Apc* mutations in the normal mucosa of *Apc*+/-*Msh2*+/-and $Apc^{1638N}Mlh1$-/- mice respectively (Sohn *et al.*, 2003) (Kuraguchi *et al.*, 2000). In Min mice inactivation of the second *Apc* allele is usually the result of LOH (Shoemaker *et al.*, 1997), instead in the *Apc*+/-*Msh2*+/- and $Apc^{1638N}Mlh1$-/- mice, a mutator phenotype caused by deficient Msh2/Mlh1 resulted in inactivation of the second allele by a somatic inactivating mutation. A high proportion of mutations observed within *Apc*+/-*Msh2*+/- adenomas were missense (56%) and this further underlines the importance of analysis into their role in tumourigenesis. In contrast, the vast majority of tumour *Apc* mutations from the $Apc^{1638N}Mlh1$-/- mouse were truncating mutations.

The DNA repair and DNA damage-induced apoptosis roles of Msh6 and Msh2 (de Wind *et al.*, 1999;Zhang *et al.*, 1999) have been successfully uncoupled using two mouse models harbouring missense mutations. The $Msh2^{G674A}$ and $Msh6^{T1217D}$ mice show loss of DNA MMR, without an affect on their apoptotic response to DNA damaging agents (Lin *et al.*, 2004;Yang *et al.*, 2004). These studies indicate that genotype-phenotype correlations can provide valuable information for cancer treatment. As tumours lacking MMR as a result of missense mutations may not have resistance to chemotherapeutic agents. Mouse models for HNPCC have been used to determine the effect of single MMR gene loss, and combinations of MMR genes on genomic mutation frequency (Hegan *et al.*, 2006). However, as mentioned previously, due to fundamental differences between humans and mice all results must be interpreted with caution. Whilst *Mlh1*-/- mice spontaneously develop gastrointestinal tumours by 48 weeks, to provide a good model for HNPCC (Tokairin

et al., 2006), mice defective in *Msh2* show increased predisposition to skin cancer (Meira *et al.*, 2002).

Mouse models have been used to look at the interactions between the Myh and Ogg1 proteins, and it is suggested that in mice they act synergistically (Xie *et al.*, 2004). A deficiency in either Myh or Ogg1 appeared to be compensated for by the presence of the other. Only in the double knockouts was the 50% survival age significantly decreased, and tumour incidence significantly increased. On the *Apc+/min* background, Sieber *et al.* show that Myh deficiency (*Myh-/-*, but not *Myh+/-*) enhances intestinal tumourigenesis, providing a model for MAP (Sieber *et al.*, 2004). Inactivating *Apc* G-T transversion mutations were observed in 10% of adenomas, however no *K-ras* mutations were observed, as is seen in human MAP. Likewise, whilst a significant increase in ACF number was observed in the *Myh-/-* mice compared with the *Myh+/-* and *Myh+/+* mice, these did not appear to progress to carcinomas. These mouse models again highlight the problems with using mice to model human conditions. Not only do Myh and Ogg1 proteins have different properties, they are also able to substitute for one another. Additionally, in the double knockout the mice do not develop CRC, but instead develop lung tumours (Xie *et al.*, 2004), and in human *MYH* mutation is not thought to predispose to lung carcinoma (Al-Tassan *et al.*, 2004).

1.10 Sequence Context of Genes, and its Role in Tumourigenesis

The *APC* mutation spectrum is known to be affected by the loss of MMR (Huang *et al.*, 1996;Otsuka *et al.*, 2003), and BER (base excision repair) (Al Tassan *et al.*, 2002;Jones *et al.*, 2002). Likewise, as discussed in section 1.7, DNA polymerases and sequence context may also play a role in influencing the type and location of mutations in *APC*.

Evidence suggests that sequence context may play a significant role in mutation susceptibility and tumourigenesis; Bacon *et al.* have shown that the sequence context of a microsatellite influences the mutation rate that is observed in MMR deficient cells (Bacon *et al.*, 2000). It was found that, when the D2S123 locus (CA repeats) was interrupted it was more stable than when the microsatellite was 'perfect'. 80% of mutations at D2S123 were derived from the shorter allele, as these contained uninterrupted CA repeats. Various oncogenes have been shown to have mutational hotspots, including *KRAS* codons 12 and 13 in CRC, but this, as with many such observed 'hot-spots' is likely to be due to ascertainment bias, as mutations at this codon seriously disrupt the functionality of the KRAS protein.

Further understanding of mutation rates and what influences them is required to be able to address how sequence context affects mutation. Neutral mutation rates are calculated either at 'four-fold' sites; for example the third codon position changes where any of the four bases could code for the same amino acid, i.e. Serine is coded for by TCT, TCC, TCA and TCG, or at ancestral repeat sequences which are areas where transposable elements were inserted and fixed before two organisms diverged. Neutral sites are often used to determine how fast a sequence is evolving, since they are not subject to the usual selection pressures. The Mouse Genome Sequencing Consortium (Waterston *et al.*, 2002) used alignments of the human and mouse genomes to study neutral mutation rates. They reported variation in the substitution rate across the human genome, that indicated "substantial regional fluctuation", which they suggest may be caused by long-range factors that affect mutation rate.

In addition to long-range factors, studies have shown sequence context effects at short ranges (Ellegren *et al.*, 2003). Silva and Kondrashov used human-baboon genomic alignments to study mutation rates, and found that nucleotide substitutions are clumped together at scales of <10, and 1000-10,000 nucleotides, indicating two different factors at work. Potentially, sequence context effects work at the shorter scale (<10bp) and regional differences in mutation rate at the longer scale (1-10Kb) (Silva and Kondrashov, 2002).

It is important to use new bioinformatics tools i.e. comparisons between complete genome sequences, to gain further understanding of mutation rates and the influence of sequence context. Such analysis will begin to highlight areas of the genome with potential mutational hot-spots, or that are mutating faster than expected. These regions can subsequently be investigated to determine if they correlate with genes and genome regions that are known to be involved in cancer.

1.11 Mutator phenotype and Genomic Instability

Genomic instability, and the role that it plays in tumourigenesis has been discussed widely in many reviews (Lengauer *et al.*, 1998;Strauss, 1998;Cahill *et al.*, 1999;Coleman and Tsongalis, 1999;Loeb, 2001;Charames and Bapat, 2003;Grady, 2004). It is largely agreed that instability is the 'engine' of tumour progression, causing no two tumours to be the same, and allowing them to adapt effectively to a changing environment. The widely held view is that an increase in mutation rate is required for the initiation and progression of cancer.

The genetic changes induced in CRC, whether heritable or sporadic can be explained by the mutator model of tumourigenesis (Loeb, 2001). This model states that mutations arising in genes that ensure the stability of the genome (i.e. MMR genes and DNA polymerases) generate mutant cells that are selected for because they are able to escape the regulatory mechanisms that control cell replication, invasion and metastasis. This model explains the enormous differences seen in the numbers of mutations in normal and cancerous cells. Stoler *et al.* used inter- (simple sequence repeat) PCR on pre-malignant polyp DNA to estimate that there are at least 11,000 genomic events per carcinoma cell in sporadic cancers (Stoler *et al.*, 1999). Although this technique only detects fairly large changes that alter the sizes of PCR products, it does suggest that genomic destabilisation is an early step, and a cause rather than effect in sporadic tumour development. There is a trade-off between the selective advantage of accumulating many mutations that allow a cell to evade constraints in growth and regulation, and the disadvantage of reaching the threshold for cell replication and viability, resulting in cell death.

The alternative to the mutator model is that pre-cancerous cells are mutating at the normal rate, estimated to be approximately 1.4×10^{-10} nucleotides/cell/division (Loeb, 2001), and that clonal expansion (or efficient selection) causes the outgrowth of those cells that randomly acquire advantageous mutations (Lengauer *et al.*, 1998). Evidence against the mutator model includes the fact that such high levels of mutation as proposed by the mutator model are likely to induce apoptosis, the fact

that tumours are dynamic and take decades to develop, and that despite being a universal process, the loss of MMR in HNPCC patients only causes a limited set of cancers (Strauss, 1998). So, heightened mutability is not necessarily accompanied by a general increase in tumourigenesis, but instead affects only a subset of specific types. In his p53 review, Strauss notes that the mutation process in some genes is different to other genes, and hypothesises that it may be likened to the generation of antibodies, where a limited portion of the genome becomes hypermutable for a limited time (Strauss, 1998).

These hypotheses indicate the enormity of trying to understand mutation rate and selection advantages, however it is important to remember that the presence of mutations within a tumour, gives no information of mutation rate, nor does it prove that a tumour is genetically unstable. It is possible that the early mutations found in human tumours provide both a selective advantage as well as an increased mutation rate (Fishel, 2001). Elucidating what is actually going on during the tumourigenic process has proven extremely difficult.

Direct examples of genetic instability include work from Bacon *et al.*(Bacon *et al.*, 2001) showing that a poly$(A)_{10}$ tract in *TGFβR2* is hypermutable in MMR deficient cells derived from a normal lineage. This leads the authors to conclude that "exon 3 of *TGFβR2* is inherently mutable and predilection for instability at (this) poly$(A)_{10}$ tract contributes to the frequent observation of *TGFβR2* mutations in MMR deficient tumours". Another example of increased mutation rate preceding tumourigenesis comes from the novel polymorphism described at codon 1307 of *APC* (Laken *et al.*, 1997). The T to A transversion at nucleotide 3920 is observed in 6% of Ashkenazi Jews and creates an A8 tract that is hypermutable, and so, indirectly causes a cancer predisposition. Approximately 50% of tumours arising in patients with this polymorphism have insertion/deletion mutations in this hypermutable run of 8 nucleotides.

The 'stable gene hypothesis' (Frei, 1990;Frei, 1995) suggests a stabilising mechanism for genes (coined as a putative 'comparase' enzyme system in the paper)

to reduce mutation rate. This hypothetical system is proposed to correct mutations occurring in any gene that appears within the genome in multiple copies (i.e. pseudogenes, and developmentally or differentially expressed copies). It follows therefore that unstable genes are likely to be those that occur in only a single copy within the genome, and are as such not protected by the comparase system. *APC* is given as an example in this paper of such an unstable gene. At present this is just an idea needing experimental evidence to support it.

Chromosomal instability is a major form of instability in the genome (detected in >85% of CRC) and arises when the cell loses its ability to ensure that the correct number of intact chromosomes are passed to each daughter cell before cytokinesis. This results in aneuploidy (a change in chromosome number) that is seen frequently in tumours. As discussed in section 1.4, CIN can be caused by truncating mutations in *APC* (Fodde *et al.*, 2001a). CIN has been detected as an early event in a significant proportion of FAP and MAP adenomas (Cardoso *et al.*, 2006). There are also an excess of 100 candidate genes that have the potential to cause a CIN phenotype (Grady, 2004), some of which may act synergistically with *APC*, further enhancing the CIN phenotype. One group of candidate genes contains the mitotic checkpoint regulators. Cahill *et al.* showed that in some colorectal cancers displaying CIN, loss of the mitotic checkpoint was associated with inactivation of the human homolog of yeast *BUB1* (*hBUB1*) (Cahill *et al.*, 1998). BUB1 controls mitotic checkpoints and chromosome segregation in yeast, by binding to the kinetochore, and is required for mitotic delay in response to spindle disruption (Roberts *et al.*, 1994;Pangilinan and Spencer, 1996). Other chromosome changes include LOH, which is caused by translocations that result in the loss of the maternal or paternal allele during recombination. This can be accompanied by the gain of the opposite allele, leaving a normal appearing karyotype, but abnormal 'allelotype' (Lengauer *et al.*, 1998). Gene amplification is another form of instability, eg. *N-myc* oncogene amplification in Neuroblastoma (Kohl *et al.*, 1983). Gene amplification is often seen to occur late in tumourigenesis, and it is likely therefore that cancer cells, unlike normal cells, may be able to survive gene amplifications (because the normal signals for apoptosis are

disrupted) making them an effect of tumourigenesis, rather than a cause in some cases.

Under normal cellular conditions the combination of repair processes and apoptosis ensures that genetic material is transmitted precisely. It has been proposed (Cahill *et al.*, 1999) that if the accumulated damage from cell division rises above the threshold for viability, cellular apoptotic pathways are activated and cell death ensues. However, if the ability for a cell to undergo apoptosis in response to DNA damage is lost, then the cell is left unchecked to accumulate mutations, and tumourigenesis is likely. Some MMR-dependent apoptosis may be mediated through MBD4, a thymine DNA glycosylase involved in reducing the mutability of methyl-CpG sites in the genome. It interacts with MLH1, and deficiency has been shown to reduce the apoptotic response following exposure to cytotoxic agents (Sansom *et al.*, 2003). Conversely, over expression of MLH1 induces an apoptotic response in both repair-proficient and -deficient cell lines (Zhang *et al.*, 1999). APC is also likely to have a role in apoptosis, i.e. through disruption of Survivin expression, as mentioned in section 1.4. There is likely to be an inverse relationship between MIN and CIN in CRC, as the two conditions are rarely found to co-exist in the same tumour. This is likely to be due to the fact that the presence of just one condition is sufficient for tumourigenesis, but may also indicate that the mutational load created by both conditions exceeds the threshold for cell replication and viability, and becomes a selective disadvantage initiating apoptosis

The specific factors that can affect the propensity to mutation, such as loss of DNA repair, have been discussed in detail already. An inherent instability is also believed to be an important influence in tumourigenesis, affecting the frequency with which one gene might accumulate mutations. The specific consequences and relative contribution of this 'inherent instability' is also not well understood. With respect to APC, we know that it plays a pivotal role in deciding cell fate. The gene also appears to be hypermutable. This is underlined by the high rate of new germline *APC* mutations that are seen in sporadic cases of FAP. 25% of all FAP cases are sporadic (new mutants) (Bisgaard *et al.*, 1994) i.e. instead of the *APC* mutation being

inherited from one or other of the parents, in this situation germline mutations occur *de novo* during oogenesis, spermatogenesis, or early embryogenesis. It should be noted however that early estimates of the proportion of cases of FAP due to new mutations is likely to be artificially elevated by MAP cases and other as yet undefined genetic conditions. Parental *APC* mutational mosaicism has been proposed to explain a proportion of the high rate of sporadic, apparently new germline *APC* mutations (Farrington and Dunlop, 1999). Farrington and Dunlop hypothesise that passage through the germline of instability-prone individuals may represent a mechanism by which new *APC* mutations might arise. However, this phenomenon has only been shown in two cases, and is unlikely to account for more than a small proportion of the sporadic *APC* mutations. Whilst being equally vital in the WNT signalling pathway, mutations in *CTNNB1* and *Axin* are seen at just a fraction of the frequency of *APC* mutations. Inactivation of *APC* is seen in virtually all CRC cases, and has been shown to be an early event.

1.12 Research aims

The central premise for this project was to investigate the hypothesis that the *APC* gene is inherently unstable. In particular to determine the influence that DNA repair plays in *APC* mutation, and how the presence or absence of DNA repair might affect mutations in individual cells and genetic alterations in whole tumours.

In view of the evidence from published literature, it was hypothesised that a number of *de novo APC* mutations and the presence of inactivating mutations in the vast majority of colorectal cancers (whether familial or sporadic) could indicate that the *APC* gene is inherently prone to mutation. The work presented in this thesis forms a body of evidence to test this hypothesis against the alternative argument that numerous mutations in the *APC* gene are seen because *APC* mutation is the most direct route to CRC. A previous investigation from the lab provided preliminary evidence that a region of the *APC* gene appeared to be hypermutable in normal cells, when compared with another gene involved in tumourigenesis (Bacon, 2001). Inherent instability is predicted to increase mutation frequency above that which is normally expected. This would explain why mutations in other vital genes such as those involved in Wnt signalling are seen at a much lower frequency than *APC*. *APC* has been shown to be susceptible to the mutagenic effect of MMR and BER deficiency, and we propose that this is also due to inherent mutability and possibly the sequence context of the gene.

I specifically aimed to investigate the inherent instability of the *APC* gene in both the *in vivo* and *in vitro* situations. Therefore, cell lines and patient samples with a variety of DNA repair process mutations were utilised. In parallel with the analysis of the underlying frequency with which mutations arise in the *APC* gene, it was deemed important to determine which of these mutations were ultimately selected during the tumourigenic process, and how selection was affected by DNA repair processes. It was predicted that evolutionary and polymorphism data would give insights into the influences that the *APC* gene is under. Alongside a detailed comparison of *de novo* mutations and mutation rates, it was aimed to use *in silico* techniques to further our

understanding of how inherent instability of *APC* might contribute to tumour progression. The final aim of the project was to look in more detail at how the loss of the DNA repair affects mutation frequency, and begin to develop other techniques that would enable this to be directly measured and compared.

Loss of APC is associated with the vast majority of colorectal cancers, and understanding the molecular basis with which these mutations arise and what influences them is clinically important in terms of identifying genetic alterations that might predict prognosis. Furthermore, as cancers with DNA repair defects influence genetic alterations, elucidation of how they exert their effect may provide new insight into the contribution of DNA repair defects in colon carcinogenesis. This may shed light on the interaction between selection and inherent instability involved in the initiation and progression of cancer.

1.13 Experimental approach

The research presented in this thesis utilises 12 different cell lines and 38 patient samples. The cell lines included lymphoblastoid, epithelial colorectal carcinoma and fibroblast cell lines that harbour defects in either MMR or BER genes, as well as wild-type cells. The patient samples were normal and tumour colonic material from individuals with sporadic colorectal cancer, or germline mutations in either MMR or BER.

We have determined the mutation frequency of *APC* using an approach where clonal selective advantage imparted by *APC* mutation was excluded. By cloning individual alleles, *APC* mutation frequency was compared between the samples, and mutation spectrum investigated.

The data presented in Chapter 3 of this thesis focuses on determining the mutation frequency and spectrum of the *APC* gene in cell lines (*in vitro*) and in patient samples (*in vivo*). The *in vitro* work compares the mutation frequency in a region of *APC* exon 15 with similar sized regions of the *TGFβR2* and *GPD1* genes. This was carried out in two WT lymphoblastoid cell lines, a MMR heterozygote lymphoblastoid cell line, 2 homozygote MMR cell lines (one lymphoblastoid and one colorectal cancer cell line), and two BER null cell lines (one fibroblast and one colorectal cancer cell line). The *TGFβR2* gene (MIM 190182) belongs to the serine-threonine kinase family of receptors. It acts by transducing the transforming growth factor-beta inhibitory signal, a signalling pathway disrupted in most colorectal cancers. As mentioned previously, (section 1.6.1) *TGFβR2* is found to be inactivated in a subset of colon cancer cell lines with MSI (Parsons *et al.*, 1995). Additionally, there is a single report of a germline mutation event in familial CRC (Lu *et al.*, 1998). *GPD1* (Glycerol 3-phosphate dehydrogenase, MIM 138420), is a house-keeping gene expressed in the colon, and used previously as a control gene in studies of MMR as it is not implicated in any disease or cancer (Geisler *et al.*, 2003). The *in vivo* work also compares mutation frequencies observed in a region of *APC* exon 15,

and *GPD1* using DNA derived from *in vivo* patient samples to deduce if selection pressures and clonal evolution might influence the mutation frequencies observed.

In Chapter 4 somatic inactivating mutations in the *APC* and *KRAS* genes are investigated in the patient samples to identify the selected mutations. The influence of DNA repair deficiency on the *APC* and *KRAS* somatic inactivating mutations could therefore be determined. The potential for using *KRAS* codon 12 mutation as a screening tool for biallelic *MUTYH* inactivation in the Scottish population is discussed.

Chapter 5 focuses on using *in silico* techniques to further corroborate the hypothesis of instability in the *APC* gene. Polymorphism and evolutionary data presented in Chapter 5 are discussed alongside the experimental results presented in Chapter 3. Here we investigated whether *APC* is subject to a high neutral mutation rate and whether purifying or negative selection is acting. The *de novo* mutation rate of *APC* is compared with a number of genes associated with other dominant traits, and the type and location of *de novo APC* mutations is compared to those arising in other similar genes.

Finally, in Chapter 6 a novel technique based on EGFP fluorescence and cell sorting has been developed for directly measuring and comparing the effect of BER deficiency on the mutation frequency in live cells. Initial stages of the optimisation of the assay are described.

A detailed discussion of the results for each section of work is provided in the relevant chapter. Chapter 7 then summarises the main themes and conclusions that have emerged during the course of this research.

Chapter 2

Materials and Methods

The following chapter documents the methods used during the course of this thesis. Where appropriate, specific details are included in the relevant chapter. Standard protocols are cited, and any adjustments have been documented. Stock solutions and media marked with an asterisk (*) were prepared by the solution and media preparation service at the MRC Human Genetics Unit, and sterilised by autoclaving. Procedures carried out by the technical services department at the MRC HGU are attributed to the relevant member of staff, or for general procedures, marked with a hash (#). Risk assessments (including COSHH regulations) were carried out for all procedures, and adhered to throughout.

2.1 Biological Material

2.1.1 Summary of cell lines

The following cell lines were used during the course of this PhD, outlined in table 2.1. Epstein-Barr virus (EBV) transformed lymphoblast cell lines ConC, ConA and lbl-c5 were established by Sheila Mcbeath (MRC, Human Genetics Unit, Edinburgh) from healthy anonymous individuals, and are maintained as laboratory stocks. Cell lines lbl-1260 and lbl-1261 were a donation from Prof. Vogelstein (John Hopkins Oncology Centre, Maryland). The colorectal carcinoma cell lines, HCT116, HCT116 + C3 and SW480 are available from American/European Type culture collections (ATCC/ECACC). The colorectal carcinoma cell line VACO425 was a donation from Dr. James Willson (Case Western Reserve University, Ohio, USA).

Prof. M. Dunlop and Dr. S. Farrington have collected a panel of samples (blood, tissue and cell lines) from patients with CRC, and these were available as a

laboratory resource. The fibroblast cell lines 2630, 439 and 825 are untransformed primary cell lines, cultured from the skin of patients with and without CRC, bearing the same MD laboratory identification number.

Table 2.1 Details of all cell lines used within this thesis, including cell type, genotype, and reference to original source.

Name	Cell Type	Genotype	Reference
ConC	LCL EBV	WT	Lab stock
ConA	LCL EBV	WT	Lab stock
Ibl-c5	LCL EBV	WT	Lab stock
Ibl-1261	LCL EBV	PMS2 deficient	Parsons *et al.*, 1995
Ibl-1260	LCL EBV	MLH1 heterozygote	Parsons *et al.*, 1995
HCT116	CRC	MLH1 deficient	Papadopoulous *et al.*,1994
HCT116 + C3	CRC	WT (restoration of MMR activity by introduction of WT allele on chromosome 3)	Koi *et al.*,1994
VACO425	CRC	MYH deficient	Parker *et al.*, 2002
SW480	CRC	WT	Leibovitz *et al.*,1976
MD2630	Fib	MYH deficient	Lab stock
MD439	Fib	WT	Lab stock
MD825	Fib	MSH2 heterozygote	Lab stock
Key:	LCL	Lymphoblastoid Cell Line	
	CRC	Epithelial Colorectal Carcinoma Cell Line	
	Fib	Fibroblast	
	EBV	Epstein-Barr virus transformed	

2.1.2 Maintenance of cell lines

The following describes the procedures used to grow and maintain all cell lines for DNA and protein extraction, transfection, apoptosis assays and FACs (fluorescence activated cell sorting) analysis:

Tissue Culture Medium

RPMI or L-15 (HCT116 only) (Gibco BRL)

10% w/v Foetal bovine serum (FBS) (Gibco BRL)

1 x Antibiotic, antimycotic solution (Sigma-Aldrich, Irvine, UK)

Phosphate Buffered Saline (PBS)*

0.1M $NaH_2PO_4.H_2O$

0.1M $NAHPO_4.7H_2O$

pH 7.4

Trypsin Versene (TV)*

50% w/v Trypsin

50% w/v Versene

Freezing Media

8% w/v Dimethyl Sulfoxide (DMSO) in FBS

Maintenance of all cell lines was carried out by myself, under sterile conditions in a category 2 containment hood in a category 2 laboratory. Long-term storage of all cell lines takes place in the liquid nitrogen facility at the MRC Human Genetics Unit. Cells are raised from liquid nitrogen by rapidly thawing in warm water, then placed into appropriate media as soon as possible, and washed twice with media, in between centrifugation steps (1200rpm in a Wifug 500E centrifuge). Cells are fed with media as required, and passaged by washing with PBS, and incubating with TV (5 min) for the adherent cell lines, and then, as for all cell lines by aliquoting 1/2 to 1/10 of cell suspension into fresh media. To maintain cell lines as a renewable stock, at least 3 x 10^6 were split from the main culture at regular intervals, centrifuged as above, and the pellet re-suspended in freezing media. Stocks are first frozen at -70°C before being transferred to liquid Nitrogen (-140°C).

2.2 DNA Purification

Solutions

1 x Tris EDTA (TE)

1M Tris (pH7.7)

0.5M Na_2EDTA (pH8.0)

3M Sodium Acetate NaOAc

3M $NaOAc\text{-}3H_2O$

pH4

2.2.1 Estimation of DNA concentration

Optical densitometry on a GeneQuant Pro RNA/DNA calculator UV spectrophotometer (Amersham Pharmacia biotech, Cambridge, UK) was used to estimate the concentration of DNA obtained from the protocols listed below. DNA samples were diluted 1 in 100 and the DNA program used to measure absorbency at 260nm and 280nm, giving a concentration in μg/ml. Stock samples were then diluted to a working stock of 100ng/μl in TE, or where starting concentration was less than this (in the case of paraffin embedded samples), to an appropriate concentration for handling.

2.2.2 Purification of DNA from bacterial plasmids

Bacterial plasmid DNA was extracted on a small scale, using a QIAprep miniprep kit (QIAGEN Ltd, Crawley, UK) according to manufacturers instructions. For DNA extraction of plasmids in a 96 well plate format a BioMek® 2000 workstation robot was used, operated by Stephen Brown (MRC, Human Genetics Unit, Edinburgh) according to manufacturers instructions.

2.2.3 Purification of DNA from paraffin embedded archival material

Prof. M. Dunlop and Dr. S. Farrington have collected a panel of blocks of tumour and adjacent normal colon mucosa samples from patients with CRC. Samples were formalin-fixed, and paraffin embedded, and are available as a laboratory resource. 10μm slides are cut from the paraffin blocks using a Shandon AS325 Retraction microtome (Anglia Scientific Instruments, Cambridge, UK), and Haemotoxylin & Eosin (H&E) stained by Naila Haq (Colon Cancer Genetics Group, Edinburgh University, UK) to identify normal and tumour tissue. Approximately 5-10 ng of tissue (normal and tumour separately) is scraped from the slides using a razor blade, and DNA extracted using a QIAamp DNA mini kit (Tissue protocol) as per manufacturers instructions (QIAGEN Ltd), with lysis performed over-night at 56^{o}C, and the final elution step carried out twice with 50μl buffer AE and 5 min incubation to insure maximum DNA yield.

2.2.4 Colorectal blood, tumour and normal mucosal DNA

The following (Table 2.2) were obtained by Prof. M. Dunlop and Dr. S Farrington as part of an ongoing collection of DNA samples from individuals with CRC. All were extracted from blood or fresh normal and tumour colon mucosa, using a QIAamp DNA mini kit, or mini blood kit, as appropriate, and tissue extracted as described above (QIAGEN Ltd).

Table 2.2 Details of all normal and tumour samples used within this thesis, including unique sample ID number, sample source and genotype details.

Sample ID	Sample Type	Genotype
1946	NM/T	Germline MLH1 carrier
2116	NM/T	Germline MLH1 carrier
533	NM/T	Germline MLH1 carrier
1974	NB/T	Germline MSH2 carrier
2486	NM/T	MSS, BER+
2490	NM/T	MSS, BER+
2491	NM/T	MSS, BER+
2494	NB/T	MSS, BER+
1184	NM/T	MYH: Germline G382D Heterozygote
1574	NM/T	MYH: Germline G382D Heterozygote
1677	NM/T	MYH: Germline G382D Heterozygote
2564	NM/T	MYH: Germline G382D Heterozygote
2596	NM/T	MYH: Germline G382D Heterozygote
3603	NM/T	MYH: Germline G382D Heterozygote
2574	NM/T	MYH: Germline G382D Homozygote
2650	NM/T	MYH: Germline G382D Homozygote
3044	NM/T	MYH: Germline G382D Homozygote
2395	NM/T	MYH: Germline G382D/Splice Site Compound Heterozygote
2256	NM/T	MYH: Germline G382D/Y165C Compound Heterozygote
2630	NM/T	MYH: Germline G382D/Y165C Compound Heterozygote
1962	NB/T	MYH: Germline Y165C Heterozygote
2009	NM/T	MYH: Germline Y165C Heterozygote
3116	NM/T	MYH: Germline Y165C Heterozygote
6121	NM/T	MYH: Germline Y165C Heterozygote
1010	NM/T	MYH: Germline Y165C Homozygote
1283	NM/T	MYH: Germline Y165C Homozygote
1299	NM/T	MYH: Germline Y165C Homozygote
1960	NB/T	MYH: Germline Y165C/P391L Compound Heterozygote
Key:	NB	Normal Blood
	NM	Normal Mucosa
	T	Tumour

2.2.5 Precipitation and clean-up of DNA

DNA was concentrated and purified using a QIAquick PCR purification kit (QIAGEN Ltd) according to manufacturers instructions, or by ethanol precipitation. For ethanol precipitation, a 1:10 volume of 3M NaOAc (pH4) was added to the DNA solution along with 2-3 volumes of 100% ethanol. The solution was incubated for 30 min at -70°C and then centrifuged for 15 min at 2000rpm in a Biofuge pico Heraeus centrifuge (Kendro Laboratory products). The DNA pellet was re-suspended in an appropriate volume of TE.

2.3 PCR Protocols

2.3.1 Oligonucleotides for PCR

Primers were supplied by Invitrogen lyophillised and re-suspended in dH_2O to a stock concentration of 1μg/μl. Working stock solutions were made to 100ng/μl with dH_2O. PCR amplification was performed on a Peltier PCT225 thermal cycler (MJ Research, Waltham USA) under the following reaction conditions, unless stated otherwise: 94°C – 3min, (94°C – 45secs, 55°C – 45secs, 72°C – 45 secs) x 35, Final extension 72°C – 10min. Primers were designed either by eye, or using the Primer3 web based design program at www://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi. Primers are listed in table 2.3, F and R refer to forward and reverse respectively.

Table 2.3 Details of all primers used within this thesis, including sequence, cycling conditions and reference where appropriate.

Primer	Sequence 5'-3'	Cycling Conditions	Reference
M13F	GTAAAACGACGGCCAG	Standard	
M13R	CAGGAAACAGCTATGAC		
APCex15B1	AAGTACAAGGATGCCAATATTATG	Standard	Groden et *al.*, 1991
APCex15B2	ACTTCTATCTTTTTCAGAACGAG		
APCex15C1	ATTTGAATACTACAGTGTTACCC		
APCex15C2	CTTGTATTCTAATTTGGCATAAGG		
APCex15D1	CTGCCCATACACATTCAAACAC		
APCex15D2	TGTTTGGGTCTTGCCCATCTTT		
APCex15E1	AGTCTTAAATATTCAGATGAGCAG		
APCex15E2	GTTTCTCTTCATTATATTTTATGCTA		
APCex15F1	AAGCCTACCAATTATAGTGAACG		
APCex15F2	AGCTGATGACAAAGATGATAATG		
APCex15G1	AAGAAACAATACAGACTTATTGTG		
APCex15G2	ATGAGTGGGGTCTCCTGAAC		
APCex15H1	ATCTCCCTCCAAAAGTGGTGC		
APCex15H2	TCCATCTGGAGTACTTTCTGTG		
APCex15I1	AGTAAATGCTGCAGTTCAGAGG		
APCex15I2	CCGTGGCATATCATCCCCC		
APCex15J1	CCCAGACTGCTTCAAAATTACC		
APCex15J2	GAGCCTCATCTGTACTTCTGC		
APCex15C1#2	ACACCAATCGACATGATGATAAT	Standard	Novel
APCex15C2#2	CAAAACTTACTTTCATCATCTTCA		
MouseAPCex15B1	AAGTATAAGGATGCCAATATCATG	Standard	Novel
MouseAPCex15D2	GCTTTGGTCTTGCCCACCTTT		
KRAScodon12F	GGCCTGCTGAAAATGACTGA	Standard	Novel
KRAScodon12R	GTCCTGCACCAGTAATATGC		
TGFBR2ex4F	TCCAACTCCTTCTCTCCTTGTT	Standard	Farrington pers.com.
TGFBR2ex4R	AGGCCAGGCTCAAGGTAAAG		
GPD1ex6F	CCTATGACCTCCACTCCTTCA	Standard	Novel
GPD1ex6R	CCCCTCTGTTCTCCCTTCTC		
GPD1ex6+7F	CTCTTCTGCAGTGGGTGTCA		
GPD1ex6+7R	CTTAATCAGTGGGGCTGTGG		

2.3.2 Standard PCR

All PCR reactions were carried out in 25μl reaction volumes, with either the Expand High Fidelity PCR system (Roche Diagnostics, Lewes, UK) or AmpliTaq (Roche Diagnostics). For the Expand High Fidelity system each reaction contained 100ng of each primer, 2.5μl dNTPs (10mM) (Invitrogen, Paisley, UK), 1x Expand buffer containing 15mM $MgCl_2$, 1 unit (0.25μl) Taq Expand, 100ng DNA and 16.75μl dH_2O. For the AmpliTaq system, each reaction contained 100ng of each primer, 2.5μl dNTPs (10mM) (Invitrogen), 1 x AmpliTaq buffer, 2.5μl $MgCl_2$ (2.5mM), 0.25μl AmpliTaq, 100ng DNA and 14.25μl dH_2O. Where PCR analysis was directly from a bacterial plasmid, colonies were stabbed into the PCR reactions using a sterile pipette tip, in place of 100ng DNA. The colony was then stabbed into a storage 96 well culture plate (see 2.7)

2.3.3 Gel Electrophoresis

Solutions

10 x Tris-Acetate EDTA(TAE)*

2M Tris

5.7% w/v Glacial acetic acid

50mM Na_2EDTA (pH8.0)

Loading Buffer

100mM Na_2EDTA (pH8.0)

0.25% w/v Bromophenol blue

30% w/v Sucrose

For resolving standard size PCR products (100-1000 base pairs) 2% agarose gels were prepared using routine grade agarose (Biogene, Kimbolton, UK) and 1x TAE, 3μl of ethidium bromide (BDH, Electran, Poole, England), per 100ml gel was added prior to pouring. 5μl PCR product was loaded into the gel with loading buffer (1μl per 4μl product). A 1Kb DNA ladder (Invitrogen,) was also loaded into each gel to determine the size of the PCR products. DNA was electrophoresed at 30-80V for

30min, or until the dye front was 1cm from the end of the gel, depending on the size of products being resolved. PCR products were visualized using a BioRad Chemi Doc system using QuantityOne (version 4.4.1) software (BioRad Laboratories, Hercules, CA, USA).

2.4 Sequence Analysis

2.4.1 Purification of PCR products

PCR products were purified for sequencing analysis using a combination of exonuclease I (USB, Ohio, USA) and shrimp alkaline phosphatase (SAP; USB). 5µl of PCR amplification mixture was added to 1µl exonuclease I (10U/µl) and 2µl SAP (1U/µl). The mixture was incubated in a Peltier PCT225 thermal cycler (MJ Research) for 15min at 37°C and then exonuclease and phosphatase inactivated by heating to 80°C for 15min.

2.4.2 DNA Sequencing

All sequencing reactions were performed using ABI PRISM Ready Big Dye Terminator cycle sequencing kit with AmpliTaq DNA polymerase FS (Applied Biosystems, Cheshire, UK).

Sequencing of purified PCR products was carried out in 10µl reactions using 100ng of purified DNA, 40ng primer (forward and reverse separately) and 4µl Big Dye Version 3.1 (Applied Biosystems). Amplification was performed in a Peltier PCT225 thermal cycler (MJ Research) under the following reaction conditions: (96°C – 30secs, 50°C – 15secs, 60°C – 4 min) x 25.

Sequencing of plasmid DNA was carried out in 22µl reactions using approximately 100ng plasmid DNA, 100ng primer, 4µl Big Dye Version 3.1 (Applied Biosystems) and 0.36mM MgCl. Amplification was performed in a Peltier PCT225 thermal cycler (MJ Research) under the following reaction conditions: 96°C – 5min, (96°C – 30secs, 50°C – 15secs, 60°C – 4 min) x 25 #.

All sequencing reactions were carried out in either eppendorf tubes or in 96 well plates.

2.4.3 Precipitation of DNA from sequencing reactions

Following amplification, precipitation of sequenced DNA was performed by adding 55µl of 95% ethanol and 2µl 3M NaOAC (pH4) to the sequencing reaction mix, and incubated at room temperature for 30 min. Samples were then spun at 1500rpm in a Biofuge pico Heraeus centrifuge for 30 min, and the DNA pellet washed in 70% ethanol, and dried in a hot-block (60°C) to allow complete evaporation of ethanol. DNA pellets were stored at -20°C prior to re-suspension.

For precipitation of DNA in a 96 well format the same procedure as above was carried out, except that plates were spun in a Sorvell RT6000 centrifuge at 2000 rpm for 30 min, and supernatant removed by flicking the plate, and then pulse spinning the upturned plates on paper towels at 800 rpm. 70% ethanol was used to wash the DNA pellets, by pipetting down the side of the wells, and rapidly inverting the plate immediately. Pellets were finally dried by a further pulse spin.

Ethanol precipitated reaction products were re-suspended in Hi-Di™ (Applied Biosystems) and heated to 90°C for 2 min by Agnes Gallacher (MRC, Human Genetics Unit, Edinburgh). Samples were resolved on the ABI PRISM® 3100 or 3730 genetic analysers, according to manufacturers instructions.

2.4.4 Analysis of sequence data

Raw sequence data was analysed using Sequencing Analysis Version 3.7 or 5.2 (for 3100 and 3730 respectively) (Applied Biosystems). The alignment of multiple sequences of the same fragment was carried out by importing all sequence data, as a project into the Consed program. Sequencing traces were read, aligned and viewed using the Phred/Phrap/Consed packages respectively (www.phrap.org/phredphrapconsed.html). The consensus sequence for each project was exported and aligned with the known sequence of each fragment using the general purpose multiple sequence alignment program ClustalW at http://www.ebi.ac.uk/clustalw, to obtain accurate mutation positions. Individual base pairs highlighted by Consed as showing mutation, deletion, insertion or ambiguous sequence were then manually curated, and mutation status confirmed as required.

2.5 Cell Analysis

2.5.1 Annexin V Apoptosis Assay

The Annexin V-FITC kit (Oncogene Research Products, San Diego, USA) was used to determine the levels of apoptosis taking place in the cell lines. As a positive control, apoptosis was induced in lymphoblastoid cells by pre-treating with 10μM Cyclohexamide (Sigma) for 1 hour, and then an overnight treatment with 10ng/ml Tumour Necrosis Factor (R&D Systems, Minneapolis, USA).

Each cell line (apoptosis induced and untreated) was split into four, and used in the Annexin V-FITC assay as described below with either FITC and Propidium Iodide (PI), FITC only, PI only, or no FITC or PI.

Cell concentration was adjusted to 1 x 10^6 cells/ml, 0.5ml cell suspension was transferred to an eppendorf, and 10μl media binding reagent (Oncogene Research Products) added. 1.25μl Annexin V-FITC (Oncogene Research Products) was added to the appropriate tubes, and incubated for 15 min at room temperature in the dark. Tubes were spun at 1300 rpm for 5 min in a Biofuge pico Heraeus, and supernatant was removed. The cell pellets were re-suspended in 0.5ml cold 1x binding buffer (Oncogene Research Products), and 10μl PI (Oncogene Research Products) added to the appropriate samples. All samples were placed on ice until analysis by flow cytometry.

2.5.2 Flow Cytometry

Assistance with Flow Cytometry was given by Kay Samuel (John Hughes Benett Laboratory, Dept. Oncology, Western General Hospital, Edinburgh), using a FACSCaliber cytometer with CellQuest software (Becton Dickinson Biosciences, Franklin Lakes, USA). The Annexin V-FITC signal was quantified at 518nm, and the Propidium Iodide signal at 620nm. The flow cytometer was set up using the

cyclohexamide/TNF apoptosis induced cells stained with FITC only, and PI only. Output data was obtained for 20,000 cells in each treatment group, and showed the log of the annexin V-FITC fluorescence on the X axis, and the log of the PI fluorescence on the Y axis

2.5.3 Transfection of cell lines

Transfection of cell lines with DNA plasmids was carried out using Lipofectamine™ 2000 (Invitrogen) in a 6 well format according to manufacturers instructions.

2.5.4 B-gal assay for transfection efficiency

A β-Galactosidase enzyme assay system (Promega, Madison, USA) was used to control for transfection efficiency in all transfection experiments. Cells were co-transfected with the pSV-β-Galactosidase control vector (Promega) and vector of interest, and β-Galactosidase activity measured in cell aliquots according to manufacturers instructions for a 96 well plate assay. Absorbance was read at 420nm using a Multiskan MS plate reader (Labsystems, Basingstoke, UK).

2.5.5 FACs analysis of GFP transfected cells

Assistance with FACs analysis was giving by Eric Miller (CR UK Cancer Research Centre, Edinburgh) using a FACSCaliber cytometer with CellQuest software (Becton Dickinson Biosciences). For the analysis of levels of green fluorescent protein expressed by transfected cell lines, cells were harvested using TV, spun at 1200 rpm in a Wifug 500E, washed, and re-suspended in 1ml PBS. The GFP signal was quantified at 507nm, and data for 5000 cells in three separate experiments (total of 15,000 cells) of each cell type, in each treatment group was acquired.

2.6 Site Directed Mutagenesis

Site directed mutagenesis was carried out on the pEGFP-C1 vector (Becton Dickinson Biosciences) using a QuikChange® site directed mutagenesis kit (Stratagene, La Jolla, USA), using the primers listed below. See chapter 5 for further details.

2.6.1 Oligonucleotides

Oligonucleotides for site directed mutagenesis were designed using a web-based primer design software program, specifically for designing optimal mutagenic primers for use with the QuikChange site directed mutagenesis kit. The program incorporates all relevant design guidelines for use with the kit, and is accessed at http://labtools.stratagene.com/QC.

The oligonucleotide sequences used are:
L45T-1 Forward 5'-TACGGCAAGCTGACCTGAAAGTTCATCTGCACC-3'
L45T-1 Reverse 5'-GGTGCAGATGAACTTTCAGGTCAGCTTGCCGTA-3'

2.7 Cloning and Bacterial Culture

Media

Luria Broth (L-Broth)*

0.1% w/v Tryptone (Difco)

0.05% w/v Yeast extract (Difco)

171mM NaCl

Luria Agar (L-Agar)*

0.1% w/v Tryptone (Difco)

0.05% w/v Yeast extract (Difco)

171mM NaCl

0.15% w/v Agar (Oxoid Ltd)

Terrific Broth (Ter-Broth)*

0.12% Tryptone (Difco)

0.24% w/v Yeast extract (Difco)

0.04% Glycerol

1 x TB phosphate

10 x TB phosphate*

0.17M KH_2PO_4

0.72M K_2HPO_4

Additives

Amplicillin Stock Solution

20mg/ml ampicillin (Sigma)

5-Bromo-4-Chloro-3-Indolyl-β-D-galactosidase (X-gal)

40mg/ml X-gal (Sigma) in Dimethylformamide (DMF)

2.7.1 TA Cloning and transformation

Individual alleles were cloned using a TOPO TA kit, version Q (Invitrogen). Genomic DNA was PCR amplified as described above (2.3.2) using the Expand High Fidelity PCR system (Roche), and purified using exonuclease I (USB) and Shrimp Alkaline Phosphatase (USB), as described above (2.4.1). Post-amplification addition of 3' deoxyadenosine (A) overhangs to the PCR products was carried out prior to ligation into the cloning vector, to ensure cloning efficiency; 1µl Taq Expand, 1µl Expand Buffer (Roche), 1µl dATP (Invitrogen) and 2µl dH_2O were added to 5µl PCR product, and incubated at 72°C for 15 min in a Peltier PCT225 thermal cycler (MJ Research). PCR products were cloned into pCR 2.1-TOPO vectors (Invitrogen) according to manufacturers instructions. Clones were then chemically transformed into TOP10 *E.coli* strains (Invitrogen) according to manufacturers instructions. 50µl of each transformation was spread onto selective L-agar plates containing ampicillin (50µg/µl) and X-gal (40µg/ml). Plates were incubated, upside down overnight at 37°C. This enabled light blue/white screening of transformants; dark blue colonies contain the vector without the insert.

2.7.2 Colony selection and storage

White and light blue bacterial colonies were picked using sterile pipette tips. Colonies were stabbed into a 96 well stock plate containing 200µl L-Broth containing ampicillin (50µg/ml), and then into the corresponding well of a 96 well PCR plate containing 24µl mastermix with M13F and M13R primers as described above (2.3.2). A PCR hot start of 94°C for 6 min was used to burst the bacterial walls prior to amplification of plasmid DNA. PCR analysis of clones indicated those that contained the correct length insert, and these were then cultured for DNA purification. Stock plates were incubated at 37°C and shaken at 160rpm overnight, long-term storage of these plates was at -70°C after the addition of 25µl glycerol to each well. DNA was purified from stock plates, by thawing on ice, and inoculation into an additional 96 well plate containing 1.2mls/well Ter-broth with 50µg/ml

ampicillin, these plates were incubated at 37°C for 18 hours shaking at 300rpm #. DNA was purified using a Biomek robot, as described above (2.2.2).

2.8 Mutation Analysis

2.8.1 MLPA

A SALSA P043 APC Multiplex Ligation-dependent Probe Amplification (MLPA) kit (MRC-Holland, Amsterdam, Netherlands) was used to quantify APC nucleic acid levels in a series of 51 unrelated individuals clinically diagnosed with a multiple polyp phenotype in whom no germline APC mutation had been identified. Blood DNA from these patients was collected and extracted by Prof. M Dunlop and Dr S. Farrington as part of an ongoing panel of samples from individuals with CRC, DNA had been extracted previously as described above (2.2.4). The MLPA technique allows DNA samples to be screened for exonic deletions and duplications that are not detectable by other mutation detection techniques, such as PCR and sequencing, as well as the most frequent small deletions at codons 1061 and 1309. 50ng of DNA from each patient, and 3 controls with known APC deletions were analysed according to manufacturers instructions using a Hybaid Touchdown thermal cycler.

2.8.2 Fragment size analysis using the ABI310 genetic analyser

Following the PCR reaction, 0.75µl of each product was mixed with 0.75µl dH_2O, 0.5µl Genescan 400HD ROX (Applied Biosystems) as an internal standard, and 12µl deionised formamide (Applied Biosystems). Products were denatured by heating to 94°C for 2 min, and then separated on an ABI310 genetic analyser (Applied Biosystems) according to manufacturers instructions, using POP-4 polymer (Applied Biosystems), 10 second injections at 15kV, and a run time of 40 min per sample at 60°C and 15kV, with filter set C.

To analyse the output data from the Genescan 350 data collection software version 1.0.4 (Applied Biosystems), peak size and areas were exported to an Excel spreadsheet, available from www.mrc-holland.com. For normalisation, relative probe signals were calculated by dividing each measured peak area by the sum of all peak

areas for that sample. Normalised peak areas are compared to the average results obtained on all samples. Samples with a reduction in relative peak area of 30% or greater were repeated to verify results.

2.9 *In Silico* Database Comparisons and Mutation Analysis

2.9.1 APC Mutation Databases

The latest version of the *APC* database, containing over 2000 mutations and the corresponding reference list, was obtained from Prof. Thierry Soussi (Laboratoire de Génotoxicologie des tumeurs, Paris) on 27th February 2004 (T. Soussi, personal communication). The database was searched to identify all germline mutations arising in FAP patients, and the corresponding references were obtained to extrapolate whether they were familial or new (*de novo*) mutations. These references are listed in Appendix C. Likewise, the Scottish *APC* database compiled by Dr Susan Farrington (S. Farrington, personal communication) was searched to identify all *de novo* germline mutations arising in FAP patients. These *de novo APC* mutations were then analysed for any patterns emerging in their spectrum (type, location etc.).

2.9.2 Mutation Analysis

The *de novo* mutation frequencies of the *APC, NF1, FGFR2, VHL, β-globin* and *α-globin* genes were obtained from published literature. Analysis and comparisons were carried out to determine how the new mutation rate of the *APC* gene compared with other genes involved in disease.

The spectrum of *de novo* mutations (2.9.1) of the *APC* gene was compared with the published database for the *PAX6* gene, and the known *de novo* mutations from the *β-globin* gene.

Polymorphism data from (Bustamante *et al.*, 2005) was obtained for the *APC* gene. The rates of synonymous and nonsynonymous substitutions were also calculated for using alignments with mouse and chimp at http://www.ensembl.org/index.html.

2.10 Protein Biology

Solutions

Lysis Buffer

50mM NaCl

10mM N-2 hydroxyethylpiperazine-N'-2-ethanesulfonic acid (HEPES) (pH8)

500mM Sucrose

1mM EDTA

0.5mM Spermidine

0.15mM Spermine

0.2% w/v Triton X-100

Hypertonic Buffer

350mM NaCl

10mM HEPES (pH8)

25% w/v Glycerol

0.1mM EDTA

0.5mM Spermidine

0.15mM Spermine

6x Sample Buffer

20% w/v Glycerol

2% w/v SDS

0.25% w/v Bromophenol blue

1 x Stacking buffer

5% w/v β-mercaptoethanol

4x Resolving Buffer

1.5M Tris

0.4% w/v SDS

pH 8.8

4x Stacking Buffer

500mM Tris

0.4% w/v SDS

pH 6.8

10x Running Buffer

250mM Tris

2M Glycine

1% w/v SDS

Semi-Dry Transfer Buffer

47mM Tris

40mM Glycine

0.037% w/v SDS

100mM Methanol

7% Resolving Gel

1x Resolving buffer

7% w/v Acrylamide

0.15% w/v Ammonium persulphate (APS)

0.01% w/v N, N, N', N', tetramethyl-1-2-diaminomethane (TEMED)

5% Stacking Gel

1x stacking buffer

5% w/v acrylamide

0.15% w/v APS

0.01% w/v TEMED

Strip Buffer

31mM Tris

0.02% w/v SDS

pH 6.7

2.10.1 Preparation of nuclear extracts

Western analysis was used to determine nuclear levels of the MMR proteins *MLH1, MSH2, MSH6*, and *PMS2*, and the BER protein *MYH* by western blotting. Cells were scraped in PBS from two confluent T25 flasks for each extract. Cells were spun at 1200 rpm in a Biofuge pico Heraeus centrifuge for 5 min, and re-suspended in 5mls PBS. Cells were then lysed in approximately 3x cell volume of lysis buffer (~80µl)

containing Complete™ protease inhibitor cocktail at 1:1250 dilution (Roche Diagnostics), 1mM pepstatin A (Sigma) and 100mM PEFA block (Roche Diagnostics). Cytoplasmic extract was removed by centrifugation at 6000rpm at 4°C for 20 min in a Micromax RF IEC centrifuge and discarding the supernatant. Nuclei were then lysed by re-suspending the pellet in approximately 1.5x cell volume (~30µl) hypertonic buffer containing Complete™ protease inhibitor cocktail at 1:1250 dilution (Roche Diagnostics), 1mM pepstatin A (Sigma) and 100mM PEFA block (Roche Diagnostics), and incubated on ice for 20 min. Debris was cleared by centrifugation at 13000rpm at 4°C for 5 min in a Micromax RF IEC centrifuge.

Total protein concentration was determined by Bradford assays (Biorad, Hercules, USA), read on a Multiskan MS plate reader (Labsystems), and nuclear extracts adjusted to 20µg by adding PBS.

2.10.2 Western analysis

Nuclear extract was added to a 1:6 dilution of sample buffer, boiled in a hot block for 5 min, and placed on ice. Samples were resolved by denaturing SDS-PAGE on a 7% polyacrylamide gel in 1 x running buffer at 100 volts for approximately 1 hour. Pre-stained molecular weight markers (Biorad) were run on each gel. Prior to transfer, gels were pre-soaked in semi-dry transfer buffer and the nitrocellulose membrane prepared by coating first in 100% methanol and then also in transfer buffer. Proteins were transferred to the nitro-cellulose membrane using a Biorad mini Trans-blot semi-dry Transfer cell for 35 min at 20 volts.

Blots were blocked in 5% dried milk (Marvel) and 0.3% Tween (Sigma) in PBS, overnight at 4°C. Blots were incubated for 1 hour with primary monoclonal antibodies (see table 2.4 below), washed in 5% dried milk (Marvel) and 0.3% Tween (Sigma) in PBS for 3 x 10 min, and then incubated for 1 hour with horseradish peroxidase conjugated sheep antirabbit IgG secondary antibody (for all primary

antibodies) (Amersham), and washed as before. β-actin was used as a loading control for all blots.

Protein bands were detected using ECL western blotting protocol (Santa Cruz Biotechnology, Buckinghamshire, UK) according to manufacturers instructions. Blots were stripped prior to re-probing with an antibody raised in the same animal by adding to a glass staining jar containing 50ml strip buffer, this was heated to 60°C using a heat block and 400µl β-mercaptoethanol added. After incubation for 40min, strip buffer was removed, blots washed in 5% dried milk (Marvel) and 0.3% Tween (Sigma) in PBS and then blocked overnight as before.

Densitometry was performed on blots to quantify protein levels. Gel pictures were scanned using a Hp Scanjet 7400c scanner and saved as .tiff files. These were used to create a volume report using QuantityOne (BioRad) software, this was opened in an Excel file, and protein band volumes normalised using the β-actin loading controls.

Table 2.4 Antibodies used in this thesis, details of manufacturer, and working concentration.

Antibody	Manufacturer	Dilution Used
MSH6	BD Biosciences	1:500
PMS2	BD Biosciences	1:500
MSH2	BD Biosciences	1:500
MLH1	BD Biosciences	1:500
MYH	Alpha Diagnostic Intl. Inc.	1:400
β-ACTIN	BD Biosciences	1:1000

2.11 Calculations and Statistics

2.11.1 Statistical Analysis

Statistical advice, in the form of confirming the most appropriate statistical test to employ for any given null hypothesis was provided by Andrew Carothers (MRC, Human Genetics Unit, Edinburgh). Chi squared (Minitab. V14) was used to determine whether the frequency of a given event was observed more often than would be expected by chance. Critical values for the Chi squared distribution were obtained from www.psychstat.smsu.edu/introbook/chisq.htm and the null hypothesis was rejected if the Chi squared statistic exceeded this value at the 95% significance level. Specific statistical calculations and methodology of data analysis is described in more detail in the relevant results chapters.

Chapter 3

Analysis of the Instability of *APC* Gene Sequences and the Influence of Defective DNA Repair

3.1 Introduction

It is apparent that the *APC* gene is mutated in the majority of colorectal cancers (Powell *et al.*, 1992;Nagase and Nakamura, 1993;Huang *et al.*, 1996;Fodde *et al.*, 2001b) however, the specific mechanisms by which these mutations are arising is not clear. One potential mechanism is that selection pressures are acting on this vital gate-keeper gene allowing advantageous mutations to become fixed. Another hypothesis is that mutations are arising spontaneously as a consequence of genomic instability. Genomic instability within multiple lineages would subsequently lead to greater clonal diversity, increasing the chances of progression to carcinogenesis (Maley *et al.*, 2006). Alternatively, an interplay between clonal selection and genomic instability is also a possibility.

Defects in DNA repair processes that maintain genomic stability provide an environment in which mutations can remain undetected, becoming fixed in subsequent cellular populations and contributing themselves to clonal expansion. Genome instability is now recognised as an important factor contributing to the initiation and progression of cancer, and in this chapter I have utilised a number of cell lines and *in vivo* patient samples to explore the hypothesis that the *APC* gene may be inherently unstable.

Our hypothesis that the *APC* gene may be inherently unstable is supported by two observations. Firstly, the rate of new germline *APC* mutations; approximately 25% of all FAP cases are thought to be due to *de novo* mutations (Bisgaard *et al.*, 1994). It should be noted however that a small number of these *de novo* mutations have been shown to be a result of mosaicism (Farrington and Dunlop, 1999), and another

unknown proportion of this 25% of cases are likely to be due to other genetic predisposing conditions, unknown at the time of the Bisgaard study, such as MAP. The second observation is the apparent susceptibility of *APC* to the mutagenic effect of DNA repair deficiencies. Lack of activity in both the MMR (Huang *et al.*, 1996) and BER (Al Tassan *et al.*, 2002;Jones *et al.*, 2002) pathways leads to hereditary colorectal cancer syndromes, in which *APC* inactivation, dysregulated Wnt signalling, and silencing of *EPHB2* expression (Batlle *et al.*, 2005) are thought to be the first pathogenic events in tumour progression.

I have employed a cloning based approach to detect the underlying genomic variation arising spontaneously during cell replication and division in individual alleles. This strategy has been used effectively before to examine the effect of defective MMR on repetitive regions of the *TGFβR2* and *BAX* genes, selective advantage imparted by mutation is excluded and therefore, unbiased reporting of true mutation frequencies is possible (Bacon *et al.*, 2001).

To address the question of whether the *APC* gene is inherently unstable I have compared the frequency of mutations arising in a specific region of *APC* (B1D2, see 3.2.1) with the frequency of mutations arising in similar sized regions from two control genes. The *TGFβR2* gene (MIM 190182) is involved in carcinogenesis, it belongs to the serine-threonine kinase family of receptors, and transduces the transforming growth factor-beta inhibitory signal; the loss of this negative regulation can contribute to tumour development. *TGFβR2* is found to be inactivated in a subset of colon cancer cell lines with MSI (Parsons *et al.*, 1995), there is one report of a germline mutation event in familial CRC (Lu *et al.*, 1998) and the signalling pathway is disrupted in most colorectal cancers. *GPD1* (Glycerol 3-phosphate dehydrogenase, MIM 138420), is a house-keeping gene expressed in the colon, and used previously as a control gene in studies of MMR as it is not implicated in any disease or cancer (Geisler *et al.*, 2003).

3.2 Methodological Overview

3.2.1 Amplification of *APC*, *TGFβR2* and *GPD1* gene fragments

A 955bp region from nt 2202 to nt 3157 upstream of the mutation cluster region in exon 15 was amplified from the *APC* gene (Acc. No. BC111591), using combinations of the primers listed in Table 2.3. A 883bp region of exon 4 of the *TGFβR2* gene from nt 65100 to nt 65983 (Acc. No. AY675319) was amplified using primers TGFBR2ex4F and TGFBR2ex4R (2.3.1). A 700bp region of the *GPD1* gene from nt 3505 to nt 4205 (ENSG00000167588), encompassing exons 6 and 7 was amplified using primers GPD1ex6+7F and GPD1ex6+7R (2.3.1). The sequence of *GPD1* intron 6 was excluded from our analysis, to prevent the possibility of different constraints on non-coding sequence affecting comparisons between the 3 gene regions. In each case 100ng DNA template was used for amplification, with the Expand High Fidelity PCR system, as described in 2.3.2.

3.2.2 Cloning of Individual *APC*, *TGFβR2* and *GPD1* alleles

Individual alleles were cloned from PCR products obtained above (3.2.1), into PCR cloning TA overhang vectors as described in 2.7.1. All clones were PCR amplified to ensure the presence of an insert of the appropriate size, by direct PCR analysis from each bacterial colony using M13F and M13R primers as described in 2.3.1. Insert positive bacterial clones were subsequently grown and plasmid DNA harvested as described in 2.7.2.

3.2.3 Sequencing of cloned alleles

Individual alleles were sequenced from plasmid DNA as described in 2.4. Each insert was sequenced directly from the positive insert clones using M13F and M13R vector

specific primers (2.3.1). Sequence data was analysed as described in 2.4.4, for the presence of mutations.

3.2.4 Statistical Analysis

The mutation frequency for each cell line was calculated at each locus analysed. The total number of mutations observed was divided by the total amount of high quality sequence analysed (in kilobases) to give a mutation frequency/Kb. It is assumed that a mutation is equally likely to occur at any individual base pair by chance. Differences in the number of nucleotides found to be mutated, out of the total number of nucleotides analysed at any given locus from different DNA templates were then evaluated for significance using a Chi-squared analysis as described in 2.11.1. The cut-off for significant p values was taken as 1%.

3.2.5 Cell lines

Epstein-Barr virus (EBV) transformed lymphoblast cell lines ConC, ConA, lbl-1260, lbl-1261 and colorectal cancer cell lines HCT116, VACO425 and the primary fibroblast cell line 2630 (2.1.1) were used. Cells were cultured as described in 2.1.2 and DNA extracted from them as described in 2.2. ConC and ConA cell lines are derived from healthy individuals with no family history of colorectal cancer. Cell line lbl-1261 is derived from patient 6 referred to by Parsons *et al.* and displays a mutator phenotype due to what was originally thought to be a dominant negative mutation in *PMS2* (Parsons *et al.*, 1995;Hamilton *et al.*, 1995;Bacon *et al.*, 2001), but has subsequently been demonstrated to be a compound heterozygote (De Vos *et al.*, 2004). Cell line lbl-1260 is derived from patient 4 (Parsons *et al.*, 1995) and displays a mutator phenotype at microsatellite sequences, but not gene sequences (Bacon *et al.*, 2001) due to an *MLH1* mutation. VACO425 has been characterised as having severely reduced levels of MYH mRNA, a complete absence of detectable MYH protein, and no 8-Oxo-G repair activity (Parker *et al.*, 2002). 2630 is a fibroblast cell

line derived from the abdominal skin of a patient with germline bi-allelic inactivation of the *MUTYH* gene.

3.2.6 Patient Samples

Normal mucosa/blood (NM/NB) and tumour (T) patient samples used in this chapter are listed in Table 2.2. The patients with sporadic CRC have been screened extensively for MMR and BER defects, and none have been found (Susan Farrington, personal communication). DNA was purified from formalin fixed, paraffin embedded samples as described in 2.2.3. DNA was purified from fresh normal and tumour colonic samples, and blood as described in 2.2.4. Due to the large numbers of patient samples analysed, the *TGFβR2* gene region was not included in the *in vivo* analysis of mutation frequencies.

3.2.7 Characterisation of cell lines by western blot analysis

Nuclear extracts were prepared from cultured cells as described in 2.10.1. Western analysis was used to determine nuclear levels of MMR and MYH proteins, as described in 2.10.2, using antibodies listed in Table 2.4.

3.2.8 Apoptosis Assay

Annexin V apoptosis assays (Oncogene Research Products, San Diego, USA) were carried out on cell lines lbl-1260, lbl-1261, ConA and ConC as described in 2.5.1. For each assay apoptosis induced cells (treated with cyclohexamide and TNF) and untreated cells were analysed for levels of early and late apoptosis. Results from the untreated cells of 3 separate assays (20,000 cells per assay) carried out on successive weeks were used to determine the basal levels of apoptosis in each cell line. Standard errors were calculated for each cell line.

3.3 Results

3.3.1 Mutation frequencies and spectra of cell lines

A detailed analysis of all mutations detected from all cell lines and genes analysed, including mutation type and location is presented in Appendix A. To address whether the *APC* gene is inherently unstable in the presence of normal DNA repair the level of spontaneous mitotic mutations arising in the B1D2 region was assessed in 2 normal (DNA repair proficient) control cell lines, ConC and ConA (Table 3.1). These frequencies were compared with those observed for ConC and ConA in two control gene regions from *TGFβR2* and *GPD1* (Table 3.1). Analysis of the observed mutation frequencies in the ConC and ConA cell lines reveals significantly higher mutation frequencies in the *APC* gene, compared to the 2 control genes *TGFβR2* and *GPD1*. ConC; $\chi 2=19.3$, $p<0.0001$, ConA; $\chi 2=14.728$, $p=0.001$. The vast majority of mutations detected in ConC and ConA at all three genes were transition mutations, with a single T:A-G:C transversion mutation detected in the *APC* gene from ConA DNA.

To further address whether the *APC* gene is inherently unstable, and understand whether mutations might arise in *APC* that are usually repaired by MMR, we have analysed the mutation frequencies of our three genes in three cell lines with MMR defects. Individual alleles from the 3 gene regions were cloned from the MMR null lymphoblastoid cell line lbl-1261, MMR null colorectal cancer cell line HCT116, and the MMR heterozygote lymphoblastoid cell line lbl-1260 (3.2.5). In contrast to the control cell lines, in all three MMR defective cell lines, no significant difference was seen between the mutation frequencies of the three genes studied (Table 3.1). However, lbl-1260 had elevated mutation frequencies at all gene sequences compared with lbl-1261 and HCT116. The mutations arising in these MMR defective cell lines displayed a similar spectrum to those observed in the control lymphoblastoid cell lines (see above). The possibility that changes in the rate of apoptosis are affecting the observed raised mutation frequency of lbl-1260 have been initially addressed using Annexin V assays (see section 2.5.1). However, neither lbl-

1260 or lbl-1261 cell lines demonstrated any increase in levels of apoptosis over control cells (data not shown).

To determine how instability in the *APC* gene might be influenced by other DNA repair deficiencies we examined two cell lines with BER deficiencies. The mutation frequencies for all genes in the 2630 and VACO425 cell lines were significantly higher than those observed in the MMR deficient cell lines, lbl-1261 and HCT116; $\chi 2=90.066$, $p<0.0001$. The *APC* frequencies of the BER null cell lines were also higher, but not significantly, than the elevated mutation frequencies observed in the normal lymphoblastoid cell lines, ConC and ConA and the MMR heterozygote cell line, lbl-1260. Overall, the propensity for transition mutations that is seen in the control and MMR deficient cell lines is also seen in the BER defective cell lines. Interestingly however, both BER null cell lines also have a number of transversion mutations present at the *APC* gene, 2 of 76 in VACO425 and 3 of 82 in 2630, representing 2.6% and 3.7% respectively. At the *GPD1* gene a single transversion mutation was identified in VACO425 DNA, representing 3.7% of total mutations. In all cell lines the spectrum of transition mutations at all genes shows an increase in the frequency of T:A-C:G and A:T-G:C changes compared with C:G-T:A and G:C-A:T changes.

Western blotting was carried out to confirm the DNA repair status of each cell line, previously inferred from the mutations harboured by the individuals from which the cell lines were derived (data not shown). The data confirmed that cell lines lbl-1260, lbl-1261 and HCT116 do indeed have MMR defects resulting from known gene mutations, and likewise that ConC and ConA are MMR proficient. Levels of MYH protein were also significantly reduced in cell lines 2630 and VACO425.

Table 3.1 Mutational analysis of *APC* exon 15 B1D2 region, *TGFβR2* exon 4 and *GPD1* exons 6 & 7 in 7 cell lines.

Gene	Cell Line	No. Clones	Total Sequence (Kb)	Number of Mutations	Mutation Frequency (Mutations/Kb)
	ConC	80	65.586	38	0.58
	ConA	71	62.159	39	0.63
	Ibl-1260	134	117.514	74	0.63
APC	Ibl-1261	154	129.68	24	0.19
	HCT116	96	85.497	20	0.23
	VACO425	91	86.663	76	0.88
	2630	94	89.864	82	0.91
	ConC	56	49.594	12	0.24
	ConA	26	25.086	3	0.12
	Ibl-1260	30	27.545	12	0.44
TGFBR2	Ibl-1261	69	67.024	10	0.15
	HCT116	41	38.014	2	0.05
	VACO425	47	39.088	24	0.61
	2630	26	22.958	19	0.83
	ConC	70	23.857	0	Undetectable
	ConA	74	25.086	5	0.2
	Ibl-1260	46	15.118	5	0.33
GPD1	Ibl-1261	20	6.674	0	Undetectable
	HCT116	26	8.602	0	Undetectable
	VACO425	109	25.397	27	1.06
	2630	69	16.077	9	0.56

3.3.2 Mutation frequencies and spectra of patient samples

To determine whether or not it is possible to detect instability of the *APC* gene *in vivo*, normal and tumour DNA's from 18 patients (3.2.6) were analysed in a similar way to the cell line DNA. Details of all mutations detected in the *in vivo* samples are listed in Appendix B.

Comparison of mutation frequencies from patients with sporadic CRC and the data obtained for *MUTYH* heterozygote and homozygote patients shows that mutation frequencies in both the *APC* and *GPD1* genes in *MUTYH* heterozygotes are generally higher than in *MUTYH* wild-type (sporadic) individuals (Figure 3.1). Additionally, *MUTYH* homozygote mutation frequencies tend to be higher than in both the sporadic and *MUTYH* heterozygotes. No significant differences were seen between the mutation frequencies detected in the *APC* and *GPD1* genes, or between normal and tumour samples.

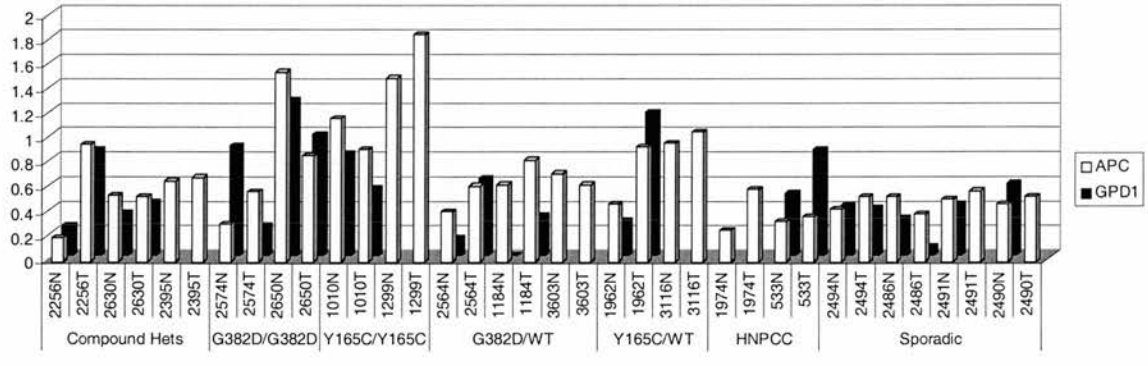

Figure 3.1 Comparison of mutation frequencies in all 36 *in vivo* samples. *MUTYH* compound heterozygotes, homozygotes and heterozygotes were compared with HNPCC and sporadic CRC patients for mutation frequencies at the *APC* and *GPD1* genes.

Overall there was an increase in transversion mutations observed in the tumour samples of *MUTYH* heterozygotes and homozygotes in contrast to the spectrums

observed in the sporadic and HNPCC individuals. No significant difference was seen in the occurrence of transversions between normal and tumour samples in sporadic and HNPCC patients. However, there was a difference in the proportion of transversion mutations present in the *APC* tumour DNA of the *MUTYH* heterozygotes, compared with the sporadic samples (p=0.014). This suggests that as a general trend there may be a heterozygous effect on mutation spectrum of having a single inactivated copy of *MUTYH*. Interestingly, G:C-T:A transversions were seen in the in the *APC* gene, but not in the *GPD1* gene in any samples studied.

3.3.3 Comparison of mutations detected in cell lines and patient samples, with known synonymous changes in *APC*

To determine whether or not the mutations detected in our cloning based assay were indeed 'real' mutations and not artefacts from the initial PCR amplification of sample DNA's, a comparison was set up between the mutations presented in 3.3.1 and 3.3.2 (and listed in Appendices A and B) and synonymous changes in the equivalent portion of the *APC* gene that have been reported in human material.

No synonymous changes are reported between codons 736 and 1054 (B1D2 portion) of the *APC* mutation database (Thierry Soussi, 2004) or the HGMD database (March 2007). A single SNP is reported in this region in the ensembl database http://www.ensembl.org/index.html. A human-mouse alignment of *APC* (see section 6.2.3) revealed 40 synonymous changes between the two sequences. Using the Xenopus gene sequence (see section 6.3.3), 14 synonymous changes were found to have occurred in the mouse lineage and 11 in the human lineage. The lineage of the remaining 15 synonymous changes could not be determined. These 11 synonymous changes, the single emsembl SNP and a single synonymous change in the human lineage detected in the human-chimp alignment (Table 6.12) were compared with the 1720 mutations detected in the B1D2 *APC* region in cell lines and patient DNA (Table 3.2 and Appendices A and B).

The synonymous mutations in Table 3.2 represent 'real' mutations, that is, mutations that have occurred naturally in the B1D2 region of *APC*, and are not under any selection pressure. If the mutations detected in cell lines and patient samples by our assay are also 'real' mutations and not PCR artefacts one would expect a comparison of the spectra of the two groups to be similar. However, despite only 13 'real' mutations being available for analysis, a highly significant difference is observed; $\chi^2=200.672$, $p<0.0001$.

				Comparison of Spectra	
Source	**nt**	**Codon**	**APC Change**	**Total**	**Assay Mutations**
Human lineage	477	894	**A-G**	**1**	**778**
Human lineage	597	934	**A-T**	**1**	**6**
Human lineage	144	783	**C-A**	**2**	**1**
SNP	409	872	C-A		
Human lineage	39	748	**C-T**	**5**	**177**
Human lineage	411	872	C-T		
Human lineage	495	900	C-T		
Human lineage	576	927	C-T		
Human lineage	726	977	C-T		
Human lineage	657	954	**G-A**	**1**	**212**
Human lineage	315	840	**T-C**	**3**	**497**
Human lineage	438	881	T-C		
Human lineage	465	890	T-C		
			Total	13	1671

Table 3.2 Synonymous changes in the B1D2 region of *APC* that have been reported in human material are compared with mutations detected in patient samples and cell lines using our cloning based assay.

3.4 Discussion

Data presented in this chapter demonstrates that a region of the *APC* gene may be hypermutable in control (MMR and BER proficient) cell lines derived from normal tissue (lymphocytes), when compared with sequence from regions in the *TGFβR2* gene, and a control gene *GPD1* (ConA, p=0.001, and ConC, p<0.0001) (3.3.1). These cells do not possess an increased mutational load from a defective DNA repair system, and a significantly high *APC* mutation frequency is revealed, whilst no such increase in mutation frequency in the *TGFβR2* or *GPD1* genes is observed. Few studies have looked at mutation frequencies in non-repetitive sequences (Takenoshita *et al.*, 1997), however, this finding is consistent with previous work showing that normal control cells can display detectable mutations (Bacon *et al.*, 2001).

This observation suggests that there may be an inherent susceptibility to mutation within the *APC* gene, which might contribute to the frequency of mutations found early in sporadic tumourigenesis. This is the first demonstration that the *APC* gene has the ability to be hypermutable in normal cells, and may also go some way to explaining why *APC* is mutated somatically in 85% of CRCs (Powell *et al.*, 1992;Miyaki *et al.*, 1994) and how certain regions of *APC*, such as the mutation cluster region are specifically affected (Miyoshi *et al.*, 1992).

The BER defective cell lines VACO425 and 2630 appear to be associated with a similar increase in mutational load as is seen in the MMR heterozygote lbl-1260 cell line, with mutation frequencies being elevated in all three genes studied (3.3.1). In the heterozygote MMR cell line lbl-1260 this may demonstrate that 'normal' cells with a heterozygote *MLH1* defect show a subtle mutator phenotype that might explain the propensity for disease progression at an earlier age. Alternatively, there may have been other events occurring in this cell line that are affecting repair that we are not aware of. When compared to the mutation frequency increase in *APC* seen in normal lymphoblastoid cells the mutation frequency increases at VACO425 and 2630 appear modest, and therefore this increase alone cannot account for the large

increase in tumour number seen in MAP patients (typically 10's to 100's adenomas compared to normal individuals of the same age with 0 to 1 adenoma).

The statistically significant increase in *APC* mutation frequencies compared with control genes in the both normal cell lines, provides evidence to support the majority of mutations detected by this assay being genuine and suggests that PCR error does not significantly bias the results obtained. However, the possibility that many of the mutations observed have been generated during PCR amplification cannot be ruled out. To go some way to addressing this issue the expected number of PCR errors was estimated for all cell lines and patient samples using the published error rate for the Expand High Fidelity PCR system (Boheringer Mannheim). These figures showed that the raised mutation frequencies observed at the *APC* gene in ConC and ConA were indeed significantly higher than expected due to Taq polymerase error ($p<0.0001$). In addition, repeat sequencing and re-isolation of plasmid stocks confirmed mutations in all cases. Nevertheless, this cannot counter the effect of mutations occurring in early clones during the culture of the cell lines (see below), during the initial PCR amplification step, or due to a property of the *APC* gene that might itself induce Taq polymerase errors (i.e. local DNA conformation). To determine whether or not PCR errors are having a significant impact on these results it would be useful to try and induce/prevent errors in the initial PCR amplification step, by using low fidelity conditions or altering DNA confirmation by cloning the sequence into a vector prior to PCR amplification. If a further increase/decrease in similar types of mutations were seen, then this would provide good evidence that these mutations are PCR induced errors and not 'real' mutations. Although this assay has been used previously (Bacon *et al.*, 2001), this study concentrated on repetitive regions where the mutation frequency is several fold higher than at non-repetitive regions, and therefore errors induced by PCR amplification made up a much less significant proportion of the errors observed.

Generally, mutation frequencies calculated from populations of cells in culture are inadequate indicators of spontaneous mutation rates. This is because the population of mutants in a culture of cells is composed of clones, each of which arose from an

individual cell that sustained a mutation. The size of a given mutant clone will depend on when during the growth of the population the mutation occurred. This is the fundamental property of spontaneous mutation that was exploited by the famous Luria Delbrück fluctuation test (Luria and Delbruck, 1943). The model was expanded in 1949 by Lea and Coulson (Lea and Coulson, 1949) to create 'The model of expansion of mutant clones'. This states that when a population is growing exponentially, as in a cell culture, the appearance of new mutants, plus the proliferation of pre-existing mutants results in a constant increase in the mutant fraction each generation. The mutation *rate* is this increase and can be determined by measuring the change in the mutant fraction over time. A population of cells must be of a sufficient size so that the probability that mutations occur each generation is the same. However, by the time the population reaches a sufficient size, some mutations will have already occurred. These will produce clones of mutants that make it impossible to accurately measure the accumulation of new mutants. To address the effect that the expansion of mutant clones might have had on the results presented in this chapter, an early passage of the lbl-1260 cell line was analysed for mutation frequency. No significant difference was detected between the early and late passages (early; 20 mutations in 39.933Kb sequence – 0.501 mutations/Kb Vs later passage; 0.63 mutations/Kb χ^2=0.83, p=0.362). Based upon the theory of expansion of mutant clones, one would expect a higher mutation frequency from the later passage, therefore this data provides further weight to the argument that the results observed in this chapter are largely due to PCR error. In addition, the data from the patient samples (3.3.2) is much more variable than that from the cell line data (3.3.1), and this is likely to be due to the vast number of highly complex selection barriers that are acting upon the *in vivo* colon during the clonal process of tumourigenesis, compared with the cell line situation (Cahill *et al.*, 1999c).

The analysis of mutation spectrums has been used previously to uncover the mechanism by which mutations are being created (Al Tassan *et al.*, 2002f), for this reason the mutation spectrums of all genes, cell lines and patient samples were examined. Previously, C:G-T:A transitions have been reported as the most common somatic *APC* point mutations in colorectal tumours (Nagase and Nakamura,

1993;Laurent-Puig *et al.*, 1998a). Here we have shown a predominance for T:A-C:G and A:T-G:C transitions, although due to small numbers significance has not been able to be tested in many cases. A slight increase in transversion mutations was detected in the VACO425 and 2630 cell lines and *MUTYH* homozygote and heterozygote patients. However, no G:C-T:A transversions were detected in the BER null cell lines by our assay, although a small number were seen in the patient samples. There were significantly more transversion mutations in the tumour DNA of the homozygote *MUTYH* patients than in the normal DNA (p=0.011). It is interesting that significantly more transversions, and specifically G:C-T:A changes have accumulated in the *APC* gene from tumour tissue compared with the normal tissue. A similar, but more highly significant effect on mutation spectrum was seen in the *MUTYH* heterozygotes, suggesting that there may be a heterozygous effect of having one inactivated copy of the MYH protein. The heterozygotes appear to show a BER defect that may not be entirely dependent on MYH, possibly implicating other BER gene involvement. Alternatively, MYH may have a role other than BER damage removal that is causing this signature of increased transversions mutations, or it could be due to ascertainment; if one assumes that the heterozygous *MUTYH* mutations are not having an affect then these patients may have other predisposing problems with DNA repair. One would expect more G:C-T:A mutations to be observed, and their absence suggests that our assay is not representative of the *in vivo* situation (oxidative damage in the colon). Infact, the overall spectrum of mutations observed, possibly with the exception of the few transversion mutations in *MUTYH* mutant patients is largely consistent with PCR derived mutations. The mutations observed in the cell line lbl-1260 and the MMR defective patient samples (1974 and 533) are also not consistent with the inactivating mutations that have been reported in *APC* in HNPCC individuals (Huang *et al.*, 1996;Huang *et al.*, 2004). These include insertions/deletions at (A)n tracts, and C:G-T:A transitions at CpG dinucleotides.

We have examined polymorphism (Bustamante *et al.*, 2005), divergence (www.ensembl.org), and chromatin environment data (J. Prendergast, personal communication) for our 3 genes of interest, and find conflicting results for which gene is predicted to be the most 'mutable'. This is discussed further in Chapter 6.

Within our data we observe similar frequencies of synonymous and non-synonymous changes occurring within the gene sequences studied, supporting the argument that what we are observing are mutations under non-selective conditions that can be used to deduce the true mutability of a sequence. In contrast to this, despite only 13 synonymous 'real' mutations being available for comparison with the 1720 mutations detected by our assay (3.3.3) we have observed that the spectrum of mutations is completely different between the two groups. This provides further evidence to support the theory that that our assay may be severely compromised by the initial PCR amplification step.

Chapter 4

The Influence of Defective DNA Repair on Somatic Inactivating Mutations in *APC* and *KRAS*

4.1 Introduction

APC inactivation is a key, early event in tumourigenesis and was thought to be necessary, and sufficient for adenoma development in all colorectal cancers (Powell *et al.*, 1992;Nagase and Nakamura, 1993). Most pathogenic *APC* mutations (germline and somatic) cause the loss or disruption of the central β-catenin binding region of the APC protein. These mutations are strongly selected for, because the subsequent nuclear translocation of β-catenin provides the cell with a selective advantage allowing for initial clonal expansion (Fodde *et al.*, 2001b). Selection for the 'second hit' in *APC* (inactivating mutation, allelic loss or promoter hypermethylation (Esteller *et al.*, 2000)) is widely believed to be dependent on the first pathogenic event (Rowan *et al.*, 2000;Albuquerque *et al.*, 2002;Cheadle *et al.*, 2002). Two truncating mutations in the MCR have been shown to be the most likely combination of events selected for, followed by one truncating mutation in the MCR with the other event either allelic loss, or a truncating mutation out with the MCR. Selection is lowest when neither event is a truncating mutation within the MCR (Cheadle *et al.*, 2002). Nevertheless, retention of one β-catenin down-regulating 20 amino acid repeat in either of the two *APC* alleles appears to be advantageous, suggesting that there is an optimal level of β-catenin signalling required for tumour formation (Albuquerque *et al.*, 2002).

In normal cells the level of cytoplasmic β-catenin is maintained at a low level by APC mediated degradation, but if APC is not functional, β-catenin is left unchecked, levels rise and the protein is free to enter the nucleus. Within the nucleus, β-catenin binds to the T-cell transcription factor 4 (TCF-4), enhancing the transcription of

many oncogenes including *c-Myc*, and cell cycle regulators such as *cyclin D1* (Bright-Thomas and Hargest, 2003). Nonetheless, whilst constitutive activation of the Wnt signalling pathway is required for progression from normal epithelium to adenoma, not all the pathway's target genes promote further neoplastic changes, and they must be consequently modified. EphB receptors are Wnt signalling target genes, despite the fact that they suppress colorectal cancer progression (Clevers and Batlle, 2006). In crypt progenitor cells they control cell compartmentalization along the crypt axis through repulsive interactions with EphrinB ligands. However, during tumour progression, when activation of β-catenin/Tcf-4 causes the initiation of tumourigenesis, selection acts against the expression of *EPHB* genes (Batlle *et al.*, 2005).

The *KRAS* oncogene codes for a small membrane attached GTP-bound protein that plays a key role in the transduction of extra cellular mitogenic signals. Mutations in *KRAS* result in the loss of GTPase activity, causing the cell to be continuously activated and locked into proliferation. Mutations are present in a significant number of different tumour types, including colorectal adenomas and carcinomas (Capella *et al.*, 1991). Between 35 and 41% of sporadic colorectal cancers are reported to harbour activating missense mutations at either *KRAS* codon 12 or 13 (Tortola *et al.*, 1999; Kambara *et al.*, 2004; Hsieh *et al.*, 2005). More recently, G:C-T:A transversion mutations at *KRAS* codon 12 have been described in the tumours of MAP patients, with variable frequencies (Lipton *et al.*, 2003; Jones *et al.*, 2004), and in *MUTYH* heterozygotes (Kambara *et al.*, 2004).

Both the *KRAS* and *APC* genes are targets for mutation in colorectal cancer, and are influenced by DNA repair defects. MAP, HNPCC and sporadic CRC patients have no inherited *APC* or *KRAS* mutations, so it is possible to compare normal and tumour material to detect the pathogenic mutations leading to a CRC phenotype. The consequence of the absence of BER or MMR on the mutation spectrum at the *APC* and *KRAS* loci can thus be elucidated. Previously, it has been shown that virtually all *APC* truncating mutations in MAP patients are G:C-T:A transversions, and almost all occur upstream of the di-nucleotide AA (Jones *et al.*, 2002; Lipton *et al.*, 2003).

However, the consequence of carrying a heterozygote *MUTYH* mutation is much less clear.

In this chapter, to further consolidate our findings from Chapter 3 we have examined the *APC* and *KRAS* genes to deduce how DNA repair affects the mutations that are selected for during carcinogenesis, and whether this tallies with the previous results. In addition, we have determined whether such mutations might be useful tools for predicting DNA repair status in Scottish colorectal cancer patients in a prognostic setting. We describe the analysis of a 2850bp region of *APC* exon 15 (codons 744-1671) that contains 8/11 of its β-catenin binding sites, and *KRAS* codons 12 and 13. Both normal and tumour tissue were amplified and sequenced, to establish somatic mutations that were present in the tumour DNA only. We have examined 4 sporadic, 3 HNPCC, 7 *MUTYH* homozygotes and 11 heterozygous *MUTYH* colorectal cancer patients from the Scottish population.

4.2 Methodological Overview

4.2.1 PCR of overlapping *APC* fragments and *KRAS* exons 12 & 13

A 2.85 Kb region of *APC* from codon 744 to codon 1671 was PCR amplified as 9 overlapping fragments BIB2 to J1J2. PCR primers are listed on Table 2.3. *KRAS* exons 12 and 13 were PCR amplified using primers KRAScodon12F and KRAScodon12R (Table 2.3). In each case ~100ng of normal and tumour DNA was amplified using AmpliTaq polymerase, as described in 2.3.2.

4.2.3 Sequencing of *APC* Fragments

PCR products were resolved on 2% agarose gels as described in 2.3.3. PCR products of the correct size were then purified using exonuclease and shrimp alkaline phosphatase as described in 2.4.1. Following purification, PCR products were sequenced using the same primers as for the amplification step, and analysed as outlined in 2.4.

4.2.4 DNA Samples

Normal mucosa/blood and tumour patient samples used in this chapter are listed in Table 2.2 (individuals 533 and 1299 were not included in the analysis in this chapter). DNA was purified from formalin fixed, paraffin embedded samples as described in 2.2.3, or from fresh normal and tumour colonic samples, and blood as described in 2.2.4.

4.2.5 Cloning of Individual alleles from sample 2650

The PCR primers APCex16G1 and APCex15G2 were used to amplify the G1I2 fragment from the tumour sample of individual 2650. This was cloned into a TA vector as described in 2.7.1/2.7.2. Sequencing of plasmid DNA was carried out as described in 2.4. Each insert was sequenced using M13F and M13R vector specific primers (2.3.1). Sequence data was analysed as described in 2.4.4 to determine if the two truncating mutations (listed on Table 5.1) occurred on the same or different alleles.

4.2.6 Statistical analysis

Differences in the number of patients with mutations in either *APC* or *KRAS* were evaluated for significance using a Chi-squared analysis as described in 2.11.1. The cut-off for significant p values was taken as 1%.

4.3 Results

Analysis of 26 archival colorectal tumours identified nine (35%) *APC* mutations and six *KRAS* mutations that are predicted to alter protein function. In none of these cases was the variant present in the constitutional DNA. Table 2.2 gives details of the genotype and samples used for the 26 patients screened for *APC* and *KRAS* mutations.

4.3.1 Somatic *KRAS* and inactivating *APC* mutations reflect bi-allelic *MUTYH* mutation status

Six of the seven (86%) individuals with biallelic inactivation of the MYH protein (*MUTYH* homozygotes), had a tumour specific G:C-T:A transversion at codon 12 of *KRAS* (Table 4.1). This was compared with none of the eighteen individuals with at least one functioning copy of *MYH* ($p<0.0001$), Table 5.2.

The MCR of *APC* was mutated in three of five (60%) *MUTYH* homozygotes for whom good quality sequencing was available for analysis. One patient demonstrated double inactivation of the *APC* gene (sample 2650). Two G:C-T:A transversions were identified in the tumour DNA of patient 2650, leading to nonsense changes at codons 1306 and 1554 (Table 4.1). Subsequent cloning, as described in 5.2.5, confirmed that these two mutations are on separate alleles (data not shown).

All detected lesions in the *MUTYH* homozygote individuals were G:C-T:A transversions, as expected to occur from unrepaired 8-oxoG lesions. All *APC* mutations occurred upstream of AA dinucleotides, and 2 are believed to be novel, i.e. they have not been previously reported in the *APC* mutation database (Feb 05 version) (Laurent-Puig *et al.*, 1998b) or HGMD (Feb 06), these are highlighted in Table 4.1.

Table 4.1 Somatic inactivating mutations detected in the tumour DNA (none of the mutations are present in constitutional DNA) of seven *MUTYH* homozygotes. Highlighted mutations are believed to be novel.

Genotype	Sample ID	APC B-J			KRAS Codons 12 & 13		
		Nucleotide Change	Amino-acid change	Sequence Context	Nucleotide Change	Amino-acid change	Sequence Context
G382D/G382D	2574	ND			34G-T	Gly12Cys	TGGTGGC
	2650	3916G-T	Glu1306X	AGAAATA	34G-T	Gly12Cys	TGGTGGC
		4660G-T	Glu1554X	AGAAAAA			
	3044	Not Screened			34G-T	Gly12Cys	TGGTGGC
G382D/Splice	2395	ND			34G-T	Gly12Cys	TGGTGGC
G382D/Y165C	2256	2821G-T	Glu941X	GGAAAAT	34G-T	Gly12Cys	TGGTGGC
Y165C/Y165C	1010	3850G-T	Glu1284X	TGAAGAT	WT		
	1283	Not Screened			34G-T	Gly12Cys	TGGTGGC

4.3.2 Somatic *KRAS* and inactivating *APC* mutations are not affected by mono-allelic *MUTYH* mutation status

In twelve *MUTYH* heterozygous individuals no *KRAS* codon 12/13 mutations were identified (Table 4.2). Four of the seven (~57%) G382D heterozygote individuals screened had inactivating tumour mutations in the mutation cluster region of *APC*. One insertion, two deletions, and one C-T transition leading to a nonsense change (Table 5.2). One of the five (20%) Y165C heterozygotes had an *APC* MCR mutation; a C:G-G:C transversion leading to a nonsense change (Table 4.2). The difference in the frequency of inactivating mutations occurring within the MCR of the *APC* gene in the two heterozygote groups was not significant (p=0.198).

One misense mutation was identified, in the constitutional DNA of patient 2596. A heterozygous C:G-T:A transition at codon 1129, leading to a Leucine to Serine amino acid change. The mutation was not present in the tumour DNA. The Leucine at this position in the *APC* gene is conserved in Chimp and Mouse, however amino acid 1129 does not fall within any of *APCs'* conserved domains. The variant amino acid has been predicted to be benign using the polyphen program, http://www.bork.embl-heidelberg.de/PolyPhen/, this program predicts the possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations. The presence of this non-pathogenic lesion in the normal, but not the tumour DNA may indicate a loss of heterozygosity event in the tumour of this patient, although this had not been pursued further.

Tumour samples from patients 1184 and 1962 both have inactivating *APC* mutations within the B1D2 region used to determine mutation frequency in Chapter 3. The mutation in 1184T occurs approximately halfway through the C1C2 fragment and the 1962T mutation at the end of the B1B2 fragment. The distribution of mutations on the individual alleles that contained either the mutant stop codon or wild-type sequence was therefore analysed, and compared with the distribution seen in the normal sample from the same patient (mutation details are listed in Appendix B). No

significant differences were observed. One might expect that the presence of an upstream stop codon would affect the frequency with which mutations occur downstream. Conceivably, this may arise due to alterations in the mechanisms controlling sequence fidelity during tumourigenesis. It is likely that due to the small size of the amplified fragments not enough sequence has been analysed to definitively determine what the effect is on downstream mutation frequency.

Interestingly, no mutations were observed at *KRAS* codon 13 in any individual screened.

In four patients with sporadic colorectal cancer (2486, 2490, 2491 and 2494), and three with HNPCC (1946, 2116 and 1974), no somatic tumour mutations were identified in our *APC* region, or at *KRAS* codon 12. The difference in the frequency of inactivating mutations occurring within the MCR of *APC* between the sporadic individuals and *MUTYH* heterozygotes was not significant (p=0.099).

Table 4.2 Somatic mutations detected in the tumour DNA (none of the mutations are present in constitutional DNA) of twelve *MUTYH* heterozygotes.

Genotype	Sample ID	APC B-J			KRAS Codons 12 & 13		
		Nucleotide Change	Amino-acid change	Sequence Context	Nucleotide Change	Amino-acid change	Sequence Context
G382D/WT	1184	2626C-T	Arg876X	G**C**GAGGT	WT		
	1677	3921-3925Del5bp	Frameshift1314X	AT**AAAAG**AA	WT		
	2564	ND			WT		
	1574	4248Del 1bp	Frameshift1418X	G**C**ATTAT	WT		
	2596	ND			WT		
	2519	4394-4395Ins2bp	Frameshift1473X	AG**(AG)**TGGAC	WT		
	3603	ND			WT		
Y165C/WT	3116	ND			WT		
	2009	ND			WT		
	1962	2510C-G	Ser837X	T**C**AAGAG	WT		
	6121	ND			WT		
	1960	ND			WT		

4.4 Discussion

The data presented in this chapter show a direct link between genetic instability and tumour gene mutations. We present data that indicate that tumour specific mutations can predict the repair status of BER null patients. In line with previous studies (Al Tassan *et al.*, 2002;Jones *et al.*, 2002), our results clearly highlight the effect that loss of BER repair has upon the somatic mutation spectrum of the *APC* gene during tumour progression. In unselected patient cohorts transitions are the most common type of point mutation leading to *APC* truncation (Luchtenborg *et al.*, 2004). Here we observe a vastly increased proportion of inactivating G:C-T:A transversions, with four of seven (57%) cancers in *MUTYH* homozygote patients harbouring an inactivating G:C-T:A mutation in the *APC* region studied. This has been seen in previous studies of CRC from BER null patients (Lipton *et al.*, 2003). No mutation hotspots were identified in the *APC* region studied, although all G:C-T:A transversions occurred at Glutamate (GAA) residues, and were therefore upstream of an AA dinucleotide.

The effect of inactivation of *MUTYH* can clearly be seen at codon 12 of the *KRAS* gene, further underlining the causal role played by genetic instability in colorectal neoplasia. Previous studies in MAP patients have shown 9 of 14 (64%) cancers and 13 of 30 (43%) adenomas to harbour *KRAS* codon 12 G:C-T:A transversions (Lipton *et al.*, 2003). Another study of 50 adenomas and 4 carcinomas from 5 MAP patients detected *KRAS* codon 12 G:C-T:A transversions in 16.7% of tumours (Jones *et al.*, 2004). Here we have observed this same mutation, resulting in a glycine-to-cysteine substitution in 6 of 7 (86%) of our *MUTYH* homozygote tumours. Jones *et al* compiled a database to determine the spectrum of *KRAS* mutations in sporadic and FAP adenomas and carcinomas. This revealed a highly significant difference in the spectrum of mutations occurring in MAP tumours (where only G:C-T:A transversions are detected) compared to the other cancer types. We have shown that the genetic instability resulting from MYH deficiency targets both the *APC* and *KRAS* genes, and whilst the *APC* changes occurred upstream of the AA dinucleotide, this is not the case at codon 12 of the *KRAS* gene. It is therefore a possibility that the

sequence context of the first G residue at codon 12 (T**G**GT) gives it susceptibility to hypermutation in the MYH deficient background, as no mutations were detected at the other 3 G residues at codons 12 or 13. Alternatively, it could be due to the nature of the mutation at codon 12 that leads it to be selected.

The question as to whether or not there is a heterozygous affect of having one inactivated copy of *MUTYH* is very much still up for debate (Kambara *et al.*, 2004;Russell *et al.*, 2005;Farrington *et al.*, 2005;Jenkins *et al.*, 2006;Tenesa *et al.*, 2006), and for this reason we have investigated twelve heterozygote individuals to determine the spectrum of inactivating *APC* and activating *KRAS* mutations. Kambara *et al* have shown a higher than expected frequency of G:C-T:A mutations at *KRAS* codons 12 and 13, and *APC* in colorectal cancer cases with mono-allelic *MUTYH* mutations, compared to those without *MUTYH* mutations (Kambara *et al.*, 2004). 5 of 20 (25%) monoallelic *MUTYH* patients Vs. 6 of 72 (8%) wild-type *MUTYH* patients had *KRAS* codon 12 or 13 G:C-T:A transversions, and 4 of 20 (20%) Vs. 5 of 72 (7%) in the MCR of *APC*, suggesting that heterozygote variants of *MUTYH* may influence genetic pathways in colorectal cancer. Our results are not in agreement with this study. We do see a subtle heterozygous effect via our mutation frequency and spectrum data (Chapter 3), but it indicates a different carcinogenic pathway compared with bi-allelic inactivation, and indeed the data in this chapter supports this different effect for mono-allelic *MUTYH* carriers at the level of somatic inactivating mutations. None of the 5 *APC* inactivating tumour mutations observed in our heterozygote patients was a G:C to T:A transversion. In the *KRAS* gene, 86% of the *MUTYH* homozygote individuals had G:C-T:A transversions at codon 12, compared with none of the 12 heterozygote individuals studied ($p<0.0001$), and none of the 4 sporadic and 3 HNPCC cases with two wild-type copies of *MUTYH*. The study by Kambara *et al* studied 20 colorectal cancers with monoallelic *MUTYH* mutations, of these only 2 had a single G382D mutation, the rest were variants of a more polymorphic nature. It is possible that the contradictory results obtained by the two studies is due to the different germline *MUTYH* mutations present in the Australian population studied by Kambara *et al.* compared to the Scottish population

studied here. Alternatively, patients classified as *MUTYH* heterozygotes by Kambara *et al.* may in fact be homozygotes, whose second mutation has not been identified.

APC and *KRAS* gene mutations are often relatively early events in colorectal tumourigenesis (Vogelstein *et al.*, 1988). However, interestingly, in this study of 16 sporadic patients (12 with single *MUTYH* mutations, and 4 with none), none (0%) had a *KRAS* mutation at either codon 12 or 13. Kambara *et al.* report finding 32 of 92 (35%), Hsieh *et al,* 36.4% and Tortola *et al* 41% of sporadic colorectal cancers to have a somatic mutation in *KRAS* at either codon 12 or 13 (Tortola *et al.*, 1999;Kambara *et al.*, 2004;Hsieh *et al.*, 2005). The most likely explanation for this phenomenon is the differences in the populations studied (Spanish, Australian and Taiwanise, respectively). We did identify 5 (33%) patients who had an *APC* mutation in the MCR, and this is line most other studies of sporadic colorectal tumours (Huang *et al.*, 1996;Diergaarde *et al.*, 2003;Hsieh *et al.*, 2005)

Previous studies have shown that the gene mutations in tumour tissue do not seem to be effective for the surveillance of colorectal cancer patients (Hsieh *et al.*, 2005). Likewise, Morrin *et al.* report that *KRAS* mutation status does not correlate with Dukes stage, tumour differentiation or five year survival rate of colorectal cancer patients (Morrin *et al.*, 1994). However, in our patient series from the Scottish population we have strong data to suggest that the presence of a G:C-T:A transversion at codon 12 of the *KRAS* gene is indicative of bi-allelic inactivation of the *MUTYH* gene, although this may not apply to other populations where heterozygotes show *KRAS* G:C-T:A mutations. As *APC* mutation is so important for colorectal cancer development, one would expect to find *APC* mutations in all the tumour samples studied here. We have analysed the entire MCR, approximately 30% of the total coding region of *APC*. It is therefore possible that there are further pathogenic mutations in other areas of the gene, deletions of the whole of, or parts of *APC* (as may be the case in patient 2596), or epigenetic alterations such as methylation (Kondo and Issa, 2004), that have not been detected by our assay. In individual 2650, both inactivating *APC* mutations were detected, and confirmed to be on separate alleles. The Glu1554X mutation occurs just downstream of the MCR,

and leaves 3 of 7 β-catenin binding repeats intact. The Glu1306X mutation is within the MCR, and causes only 1 β-catenin binding repeat to remain intact. Although two mutations within the MCR are thought to be the most likely combination of events leading to *APC* inactivation, one within and one out with is also highly probable (Cheadle *et al.*, 2002). Additionally, as has been predicted previously, some β-catenin binding activity is likely to remain, despite these inactivating mutations (Albuquerque *et al.*, 2002).

In this chapter, detailed analysis of a series of tumours has allowed independent determination of the effect of genetic instability on mutations in two genes known to be involved in tumourigenesis. We highlight that as well as increasing the overall genomic mutational load of a cell, the presence of inactivated BER leads to an increase in the proportion of total transversions (somatic) mutations in the colon. The susceptibility of the *APC* and *KRAS* genes to such mutations is likely due to the presence of specific target sequences, such as at *KRAS* codon 12, and this goes some way to explain why CRC forms part of the MAP tumour spectrum.

Chapter 5

In Silico Analysis of Genome Mutation Frequencies

5.1 Introduction

In the chapter I have set out to investigate the possibility of instability in the *APC* gene sequence using an alternative approach, including number of *in silico* techniques, to that presented in Chapter 3.

Instability of the *APC* gene within the genome might explain the high frequency of new mutations observed in FAP, (these *de novo* mutations arise without the usual pattern of autosomal dominant inheritance, instead germline mutations occur *de novo* during oogenesis, spermatogenesis, or early embryogenesis), and the fact that the majority of colorectal cancers (whether familial, or sporadic) carry *APC* mutations. The idea of unstable genes (Frei, 1990;Frei, 1995) and hypermutable regions of the genome (Strauss, 1998) have been proposed before. It is widely accepted that genomic instability is required for the initiation and progression of cancer (Cahill *et al.*, 1999). It has recently been proposed that cellular and clonal diversity are key parameters driving neoplastic transformation and progression (Maley *et al.*, 2006), with multi-lineage persistence, rather than sequential selection favouring increased diversity (Shibata, 2006).

In order to address the *APC* genomic mutation status on a wider basis, I have compared the *de novo* mutation rate of *APC* and other disease genes for which data is available. This will allow me to determine if the published mutation rate for *APC* is indeed higher than for other similarly inherited, similarly sized genes. Using disease mutation databases I have established where, and by what mechanisms *de novo APC* mutations are likely to have occurred from, and how this compares to familial *APC* mutations and *de novo* mutations in other genes. Mutations in the *PAX6* gene have previously been compared to those in the *Factor IX* gene to ascertain what occurs at

the mutational level across disease genes (Prosser and vanHeyningen, 1998), highlighting this as a meaningful and valid technique. From the pattern of mutations within genes, previous studies have been successful in inferring mechanisms by which they arise (Kuraguchi *et al.*, 2000;Al Tassan *et al.*, 2002;Shao *et al.*, 2004), therefore knowledge of the type and location of *de novo APC* mutations may be insightful. Using orthologs I have also determined the degree of conservation of the *APC* gene in mouse and chimp, and the degree of polymorphism in *APC* amongst the general population. If *APC* is indeed highly mutable, as the previous experiments have suggested, then one might expect it to have a reduced degree of evolutionary conservation, and harbour a larger number of silent polymorphisms compared to other similar genes, unless there is some constraint mechanism in place for maintaining the DNA sequence of *APC* constant.

5.2 Methodological Overview

5.2.1 Comparison of *de novo* mutation rate in disease genes

As outlined in 2.9.2, Entrez was used to search the Pubmed database (www.pubmed.gov) to access relevant citations from biomedical literature that had calculated the *de novo* mutation rate for an autosomal dominant disease gene. Original articles were studied to ensure that mutation rates had been calculated in comparable ways, by the direct method (Vogul and Motulsky, 1986):

$$\text{Mutation Rate } (\mu) = \frac{\text{Number of sporadic } \textit{(de novo)} \text{ cases with condition of interest}}{2 \times (\text{Number of individuals examined})}$$

This calculation gives the probability with which a particular mutational event takes place per generation. This is done by determining the number of new alleles carrying a *de novo* mutation within a certain period of time (i.e. individuals born with FAP in a 30 year period) as a proportion of the total number of all new alleles (2 x the number of individuals born in same 30 year period). The direct method can be used for single gene determined traits as well as genomic and chromosome mutations, but cannot be used for recessive conditions.

5.2.2 Investigation of the type and position of *de novo* mutations in disease genes

To address how and where *de novo APC* mutations occur the latest version of the *APC* mutation database (Beroud and Soussi, 1996) was obtained from Thierry Soussi (personal communication) and manipulated, as outlined in 2.9.1.

2020 *APC* mutations were listed in the original database, 1000 of which were classified as 'germline FAP mutations'. The germline FAP mutations were further classified by referring to the original references cited in the database, and listed in

Appendix C. A total of 166 original references were analysed to determine the status of the mutations to which they were referring as 'true familial' (shown to have been inherited), sporadic (*de novo* mutation), or unknown (if no data on mutation origin was available from the original reference). This information was then added to the database. References containing details of *de novo* mutations are highlighted in Appendix C.

In total 708 mutations were classified as 'true familial', and 38 as sporadic/*de novo*, for the remaining 1274 no data was available on mutation origin. The Scottish *APC* database compiled by Dr Susan Farrington (S. Farrington, personal communication) was organised in a similar manner. An additional 56 Scottish sporadic/*de novo* cases were added to the 38 from the Soussi database. The total of 94 sporadic/*de novo* germline *APC* mutations (Appendix C) were then compared with the 708 true familial germline mutations for mutation position, wild-type and mutant codon/amino acid, event (i.e. introduction of a stop codon) and mutation type (transition, transversion, frameshift).

The sporadic/*de novo APC* mutations were also compared with sporadic/*de novo* mutations in the *PAX6* and *β-globin* genes. The *PAX6* mutation database (http://pax6.hgu.mrc.ac.uk/) was searched for sporadic mutations, and information was compiled into a similar format as the *APC* database (Appendix C). The 40 mutations listed in the Stamatoyannopoulos and Nute paper on *de novo* mutation rate in *β-globin* (Table 6.2), along with 10 *β-globin* mutations leading to Hb M disease were also compiled into database format (Appendix C) (Stamatoyannopoulos and Nute, 1982). The three sporadic/*de novo* mutation databases were used to perform a three-way comparison of sporadic germline mutations (mutation type, WT/mutant codon and mutation position).

5.2.3 Sequence conservation and polymorphism in *APC*

Gene sequences and polymorphism data were obtained from the ensembl genome browser (www.ensembl.org). Gene ID's for the *APC* gene in human (*Homo sapien*), chimpanzee (*Pan troglodytes*) and mouse (*Mus musculus*) are listed in Table 6.1.

Table 5.1 Ensembl gene IDs for *APC*

Species	Ensembl Gene ID
Homo sapien	ENSG00000134982
Pan troglodytes	ENSPTRG00000017135
Mus musculus	ENSMUSG00000005871

Pairwise alignments were performed using the 'Needlman & Wunsch' algorithm at the former UK human genome mapping project resource centre (www.hgmp.mrc.ac.uk).

The phylogenetic analysis by maximum likelihood (PAML) resource at (http://abacus.gene.ucl.ac.uk/software/paml.html) was used for the estimation of synonymous and nonsynonymous substitution rates for both the human-chimp and human-mouse alignments using the codonml program. This is a maximum likelihood analysis of protein-coding DNA sequences using codon substitution models (Goldman and Yang, 1994) for the calculation of codon-usage tables, the estimation of synonymous and nonsynonymous substitution rates, likelihood ratio testing of positive selection or relaxed selective constraints along lineages based on the d_N/d_S rate ratios, identification of amino acid sites or evolutionary lineages potentially under positive selection, and the reconstruction of ancestral codon sequences.

5.3 Results

5.3.1 Published *de novo* mutation rates in *APC* and other genes

Every individual born in Denmark with FAP between 1920 to 1949 has been identified and recruited to the Danish polyposis register (Bulow, 1991). Probands and family members have been analysed to deduce whether the *APC* mutation is familial or sporadic. For the same period (1920 to 1949) the number of births was obtained from Vital Statistics, Denmark (http://www.dst.dk/HomeUK.aspx). The *de novo* mutation rate of *APC* was calculated by the direct method (5.2.1) to be 9.3 x 10^{-6} mutations/genome/generation (Bisgaard *et al.*, 1994). In this section this rate has been compared to the published *de novo* mutation rates of seven other autosomal dominant conditions and their responsible genes.

Table 5.2 lists the mutation rate estimates for the seven conditions including FAP. Inclusion in the table depends upon the reliable determination of true *de novo* cases, and population size/birth rate in the same period from which cases were determined. The following is a brief synopsis of the conditions/genes included in Table 5.2.

Neurofibromatosis (MIM 162200), one of the most common inherited diseases known to man, is caused by mutations in the tumour suppressor gene *Neurofibromin* (*NF1*, 17q11.2); 4.29 x 10^{-5} mutations/genome/generation. It is a large gene with a predominant paternal derivation of mutations (Stephens *et al.*, 1992). The condition is characterised by café-au-lait spots, and fibromatous tumours of the skin.

Prader-Willi Syndrome (MIM 176270) is caused by genomic deletions at 15q11.2-q12 in the male germline (maternal genes at this locus are inactive due to imprinting). The vast majority of cases are *de novo* interstitial deletion, 2.86 x 10^{-5} mutations/genome/generation. The condition is characterised by diminished foetal activity, obesity, muscular hypotonia, mental retardation, short stature and hypogonadotrophic hypogonadism (Prader, 1956).

Apert Syndrome (MIM 101200) is caused by mutations in the *Fibroblast Growth Factor Receptor 2* gene (*FGFR2*, 10q26.13); 7.75 x 10^{-6} mutations/genome/generation. Most cases are believed to be sporadic, and a paternal age effect has been reported (Moloney *et al.*, 1996). It is a complex syndrome consisting of skull malformations, complete distal fusion of the fingers and a tendency for fusion of the bones.

Von Hippel-Lindau Disease (MIM 193300) is caused by mutations in the tumour suppressor gene *VHL* (3p25.3); 5.15 x 10^{-6} mutations/genome/generation. It is a cancer syndrome, predisposing to a variety of malignant and benign neoplasms, most frequently retinal, cerebellar, and spinal hemangioblastoma, renal cell carcinoma, pheochromocytoma and pancreatic tumours.

Methemoglobinemic Cyanosis (HbM) (MIM 141800) (16p13.3) is caused by substitution of critical Histidine residues of the α-globin chain with Tyrosine, 4.3 x 10^{-8} mutations/genome/generation. HbM causes cyanosis from birth (Stamatoyannopoulos and Nute, 1982).

Unstable Hb Disease (MIM 141900) (11p15.5) is caused by structural abnormalities that lead to the introduction of an abnormal charge to the interior of the β-globin chain, alterations between subunits, or distortion of the α-helical configuration of the affected chain (Stamatoyannopoulos and Nute, 1982), 1.32 x 10^{-8} mutations/genome/generation.

Comparison of the data sets presented in Table 5.2 show that the rate of *de novo* mutations arising in the *APC* gene and leading to FAP, is 4 fold lower than in the *NF1* gene, and 3 fold lower than at the chromosome 15 region associated with Prader-Willi syndrome. In contrast, the new mutation rate of *APC* is 1.2 fold greater than seen in the *FGFR2* gene, 1.8 fold greater than seen in the *VHL* gene, and 234 and 936 fold greater than seen in the *β-* and *α-globin* genes respectively.

Differences in estimated mutation rates which are deduced from clinical phenotype presentation may be the result of factors other than the presence of an underlying difference in germline mutation rate. Most of these will be considered in detail in the discussion section of this chapter. However, differences in the number of nucleotides per gene will be examined here. Table 5.3 shows the mutation rates for the seven genes of interest, corrected for gene size (number of coding base-pairs). By correcting for gene size, there has been a considerable change in the order of the mutation rates with respect to the *APC* gene. The *VHL* gene now has a mutation rate over 7 fold greater than *APC*. Due to its large size, the *NF1* gene is no longer the gene with the highest mutation rate. The mutation rate of *NF1* is still around 4 fold greater than *APC*, due to the fact that both genes are a similar size (8517 and 8535 coding base-pairs respectively). The mutation rate of *FGFR2* is 2 fold greater than *APC* when considered per base pair, rather than 1.2 fold less before the correction was made, however it should be noted that Apert syndrome is caused by specific missense substitutions involving just two amino acids of *FGFR2* that are selected for in the male germline. For this reason the mutation rate of *FGFR2* cannot strictly be compared with that of *APC* in this way. Due to the small size of both the α- and β-*globin* genes (438 and 423 base-pairs respectively) the fold difference between their mutation rates and that of *APC* has been reduced dramatically to 11 fold less for β-*globin* and 35 fold less for α-*globin*. The mutation rate determined for Prader-Willi syndrome cannot be accurately adapted to give a per base pair figure. Whilst the majority of cases are *de novo*, they arise as a result of interstitial deletion affecting a critical region of chromosome 15, rather than changes in a specific gene. It is reasonable to assume that when considering mutation rate corrected for gene size this would result in a relatively low mutation rate, as the critical region is so large. These corrections take into account gene size, but not other differences that might influence differences in estimated mutation rate such as sequence specificity, or the influence of disease biology (see discussion). Thus, with correction for gene size, *APC* is in the middle order section when considering *de novo* mutation rate.

Table 5.2 Estimates of mutation rates (calculated by the direct method) for 7 autosomal dominant conditions, including FAP. Data from which rates were calculated; number of new mutants (*de novo* cases) and number of births in same period are included. Mutation rates are expressed as the number of mutations per million gametes per generation for easy comparison between genes. Genes are listed in order of mutation rate, with the highest rate appearing at the top of the table.

Gene	Condition	Ref	Number of New Mutants	Number of births in same period	Number of mutations per gamete (genome) per generation	Number of Mutations per million gametes per generation
NF1	Neurofibromatosis	Samuelsson and Akeson	29	337,979	4.29E-05	42.90
15q11-13	Prader-Willi syndrome	Hiroaki *et al.*	15	262,000	2.86E-05	28.63
APC	FAP	Bisgaard *et al.*	39	2,083,366	9.36E-06	9.36
FGFR2	Apert Syndrome	Cohen *et al.*	57	3,677,419	7.75E-06	7.75
VHL	von Hippel-Lindau disease	Maher *et al.*	5	485,437	5.15E-06	5.15
β-Globin	Unstable Hb Disease	Stamatoyannopoulos and Nute	40	464,800,000	4.30E-08	0.04
α-Globin	Hbs M	Stamatoyannopoulos and Nute	5	189,700,000	1.32E-08	0.01

Table 5.3 Estimates of mutation rates (calculated by the direct method) for 7 autosomal dominant conditions corrected for gene size, therefore representing a per-base-pair, per-generation rate. Genes are listed in order of mutation rate, with the highest rate appearing at the top of the table.

Gene	Condition	Number of mutations per base pair per generation
VHL	von Hippel-Lindau disease	8.06E-09
NF1	Neurofibromatosis	5.04E-09
FGFR2	Apert Syndrome	3.15E-09
APC	FAP	1.10E-09
β-Globin	Unstable Hb Disease	9.82E-11
α-Globin	Hbs M	3.12E-11
15q11-13	Prader-Willi syndrome	Large chromosome region

5.3.2 Analysis of the type and position of *de novo* mutations in disease genes

Few studies have focused on the genetic features of sporadic FAP. In a comparison of the clinical and genetic characteristics of patients with and without a family history of FAP, Truta *et al.* have shown that truncating *APC* mutations are detected less frequently in patients without a family history (Truta *et al.*, 2005), however, screening was not carried out for genomic deletions/re-arrangements. We have comprehensively screened 51 unrelated individuals, clinically diagnosed with a multiple polyp phenotype, and without a family history of disease, in whom no germline *APC* mutations had been identified, for exonic deletions and duplications by MLPA, as described in 2.8.1 and 2.8.2. Two deletions were identified (data not shown), accounting for 4% of cases. An unknown proportion of *APC* mutation negative polyposis patients are expected to carry mutations in recessive genes such as *MYH*. MAP, as discussed in Chapter 1, can predispose to a similar phenotype to that seen in FAP. However, in the Scottish population MYH deficiency appears to play very little role in classical FAP (Susan Farrington, personal communication). As a recessive condition, MAP may occur in individuals without a previous family history of colorectal cancer, and therefore appear sporadic. There are also likely to be other, as yet undiscovered genes involved in colorectal cancer that may explain a proportion of these cases. For the remaining cases it has been proposed by Truta *et al.* that the type of *APC* mutation may differ between FAP cases with and without a family history. If this is indeed correct, they propose that sporadic FAP could be classified as a distinct condition. The techniques used by the Truta *et al.* study to detect mutations in FAP (SSCP and PTT) focus upon protein truncating mutations. Intronic and missense mutations are not readily detected using these techniques. It has been estimated that germline *APC* missense variants such as E1317Q account for ~10% of patients with multiple adenomas (Lamlum *et al.*, 2000a). An enrichment of sporadic FAP cases with these types of mutation (with potentially low penetrance) could explain why patients without a family history are more often classified as mutation negative. Data from this laboratory has revealed that 1.12% of controls also

display heterozygosity at this specific locus (Susan Farrington, personal communication).

To approach this problem we have used 2 databases of *APC* mutations causing FAP to carry out a direct comparison of *de novo* and familial mutations. These databases are collections of data from many separate studies that have used a variety of techniques to detect the mutations.

The majority of *APC* mutations in both the *de novo* and familial groups were located in exon 15; 71/91 (78%) of *de novo* mutations, 520/708 (73%) of familial mutations. The remaining mutations were evenly distributed between exons 4 to 14. No significant differences were found between the two groups of *APC* mutations when examined for mutation type, or mutated and mutant codons. Data are shown in Tables 5.4, 5.5 and 5.6.

Table 5.4 Distribution of mutation type is similar between the *de novo* and familial *APC* mutations.

	de novo		Familial	
	n		n	
Transition (Ts)	17	18%	138	19%
Transversion (Tv)	8	9%	63	9%
Frameshift	65	69%	507	72%
Splice Site	1	1%	0	0%
Large deletion	3	3%	0	0%

Table 5.5 Distribution of mutated codon types, shown as percentages, no significant differences are observed between the *de novo* and familial *APC* mutations.

	de novo (%)	Familial (%)
Arg	20	23
Asn	3.5	2.5
Gln	11	11
Glu	26	28
Gly	1	1.5
Ile	2	3
Lys	12	16
Pro	1	2
Ser	7	9
Tyr	2	4
Ala	2	0
His	1	0
Leu	4	0
Phe	1	0
Val	3	0
Asp	3.5	0
Cys	0	0

Table 5.6 Distribution of mutant stop codon types is similar between the *de novo* and familial *APC* mutations.

	de novo		Familial	
	n		n	
TAA	5	5.5%	27	4%
TAG	5	5.5%	51	7%
TGA	15	16%	109	15%
Frameshift	69	73%	521	74%

Although no differences are observed between *APC* familial and sporadic mutations, the distribution of *de novo* mutation types differs significantly between the *APC* and the *β-globin* genes (Table 5.7). In the *β-globin* gene all sporadic disease causing mutations are missense. In *APC* 69% of mutations are small insertions or deletions that cause a frameshift leading to early termination of the protein (Table 5.7). This difference reflects the fact that missense mutations in *APC* are largely unreported, as their phenotype is difficult to determine (Luchtenborg *et al.*, 2004). In unstable Hb, and Hb M disease missense mutations cause the disease phenotype (Stamatoyannopoulos and Nute, 1982). However, Stamatoyannopoulos and Nute looked for all types of haemoglobin disorders and all but one of the *de novo* cases caused unstable Hb disease, or Hb M disease, therefore virtually all *de novo* mutations in this gene *are* missense mutations. A scan of literature published since this initial 1982 report shows that subsequently reported *de novo* haemoglobin disease mutations are also virtually all missense. (search using keywords '*de novo* haemoglobin disease' at www.pubmed.gov)

In *PAX6* a mixture of missense, nonsense and frameshift mutations are reported with additional splicing mutations also disrupting normal protein function (Table 5.7). The lack of reported *de novo* splicing mutations in *APC* and *β-globin* is likely to be due to reduced analysis of intronic sequences, and in the case of *β-globin* reflect the fact that it is a much smaller gene (Figure 5.1), with just 3 exons.

Sporadic transition mutations at CpG dinucleotides are presumed to result from spontaneous deamination of endogenously methylated 5-methylcytosine (Kondo and Issa, 2004). The data presented here show a significant difference in the proportion of *de novo* transition mutations that occur at CpG dinucleotides in our three genes of interest (χ^2=21.443, p<0.0001), due to the distribution of CpG and non-CpG transitions in the *β-globin* gene (Table 5.7). 86% of mutations arising in the *β-globin* gene occur at codons coding for Histidine, Leucine, Phenylalanine or Valine. None of which contains a CpG. Previous reports demonstrate that 28% of all *PAX6* mutations are C-T transitions at CpG dinucleotides (Prosser and vanHeyningen, 1998), this analysis has indicated that *de novo PAX6* mutations alone show a slight

decrease in transitions at CpG dinucleotides to 20% (i.e. 22 of the total 112 *PAX6* mutations are transitions at CpG) (Table 5.7). Again, no significant difference between familial and *de novo* transitions at CpG was observed for *APC*; of the 138 transitions seen in familial mutations (Table 5.4) 89 (64%) occurred at a CpG dinucleotide, compared with 13/17 (76%) of *de novo* mutations (Table 5.7). All of these mutations (familial and *de novo*) occur at CGA Arginine codons, and via a C-T mutation result in TGA stop codons (Table 5.8). The high rate of *de novo* mutations in *APC* cannot therefore be explained by excessive deamination of 5-methylcytosine residues, as these only make up 14% of all mutations. However, a significant proportion of all truncating mutations in *APC* are likely to have arisen originally as a result of transition mutations at CpG. In all three genes approximately two thirds of all sporadic nucleotide substitutions are transitions (Table 5.8).

Taking into account ascertainment bias, and the over-reporting of certain mutations, comparisons with *PAX6* and β–globin *de novo* mutations show that *APC* does not appear to have an unusual pattern of mutation types that might explain the high rate of *de novo* mutations.

Table 5.7 Types of *de novo* mutations in three genes of interest

	β-globin		*PAX6*		*APC*	
	n		n		n	
Ts total	**35**	**70%**	**41**	**36.5%**	**17**	**18%**
Ts @ CpG	5	14%	22	54%	13	76%
Ts @ non-CpG	30	86%	19	46%	4	24%
Tv total	**15**	**30%**	**18**	**16%**	**8**	**9%**
Tv @ CpG	0	0%	6	33%	1	12.5%
Tv @ non-CpG	15	100%	12	67%	7	87.5%
Frameshift	**0**	**0%**	**30**	**27%**	**65**	**69%**
In frame insertion/deletion	**0**	**0%**	**2**	**2%**	**0**	**0%**
Splicing mutation	**0**	**0%**	**21**	**18.5%**	**1**	**1%**
Large deletion	**0**	**0%**	**0**	**0%**	**3**	**3%**
Total	**50**		**112**		**94**	

Table 5.8 Proportions of Transitions and Transversions in three genes of interest

	β-globin		*PAX6*		*APC*	
	n		n		n	
Transitions						
A->G	1	2%	8	14%	0	0%
C->T	9	18%	28	47%	17	68%
G->A	5	10%	2	3%	0	0%
T->C	20	40%	3	5%	0	0%
Transversions						
A->C	1	2%	0	0%	0	0%
A->T	0	0%	8	14%	1	4%
C->A	3	6%	3	5%	3	12%
C->G	0	0%	1	2%	1	4%
G->C	2	4%	0	0%	0	0%
G->T	1	2%	4	7%	3	12%
T->A	4	8%	0	0%	0	0%
T->G	4	8%	2	3%	0	0%
Total	**50**		**59**		**25**	

As mentioned previously, in *β-globin* 60% of mutated codons occur at either Histidine or Leucine, with the other 40% coding for just 5 other amino acids. This can be explained partly by the biology of methemoglobinemic cyansosis. HbM disease is usually produced by the substitution of a proximal or distal Histidine by a Tyrosine. These Histidine residues are critical for the binding of the heme group, which is disrupted by changes to the charge of the α-globin chain when replaced by Tyrosine (Ameri *et al.*, 1999). This accounts for 10 of the 11 Histidine mutated codons. *PAX6* has the widest variety of mutated codons (only Histidine codons remain unmutated in sporadic cases) although in 26% of mutations, Arginine is the target, and in 13% the target is Glutamine. In *APC* the most frequently mutated codons in sporadic cases are those most easily mutated to a stop codon with just a single transition or transversion. Table 5.10 shows that all substitution mutations lead to a stop codon. Selection is therefore likely to be the most significant factor affecting the distribution of mutated codons, with the FAP phenotype manifesting via truncating mutations.

Figure 5.1 shows that the APC protein contains an armadillo repeat region at its 5' end, and a microtubule binding/EB1 binding domain at the 3' end. The β-catenin regulatory regions are found in the central portion of the gene (containing the mutation cluster region, amino acids 1286-1513). Armadillo repeats consist of approximately 40 amino acid repeats, forming a super-helix of helices that is proposed to mediate interaction with β-catenin. From the *de novo APC* mutations, 2 previously reported mutation hot-spots are again detected, at codon 1061 (10% *de novo* mutations) and codon 1309 (19% *de novo* mutations) (Table 5.11). 31% of *de novo* mutations are found to lie within the MCR, this does not differ significantly from the 29% of germline mutations reported previously (Miyaki *et al.*, 1995).

PAX6 contains a conserved 124 amino acid N-terminal paired box domain (residues 42 to 180) of unknown function. The C terminal homeobox domain binds DNA through a helix-turn-helix (HTH) structure and spans the residues 263 to 319. The new mutations recorded in the *PAX6* database appear fairly uniform throughout the gene and are not clustered disproportionally in the conserved domains. Clustering of

mutations in the conserved domains has been reported previously in a study that did not separate *de novo* and familial mutations (Prosser and vanHeyningen, 1998). 33% of the gene constitutes the paired box domain, and 28% of new mutations are found here, likewise, the homeobox domain covers 13% of the gene, and 17% of mutations are found within it. The majority of mutations in the homeobox domain occur at a CpG dinucleotide at codon 240. This is the most frequent *PAX6* mutation, accounting for 12% of all reported mutations (Prosser and vanHeyningen, 1998), but only 6% (7/113) of *de novo* mutations (Table 5.11). The β-globin protein contains a globin domain from residues 7 to 141 (over 90% of the protein) within which, all but 1 *de novo* mutations are contained (Table 5.11).

Table 5.9 Codons that are targets for mutation in the three genes of interest

	β-globin		*PAX6*		*APC*	
	n		n		n	
Arg	0	0%	24	27%	17	19%
Asn	0	0%	4	4.5%	4	4.5%
Gln	0	0%	12	13%	11	12.5%
Glu	0	0%	4	4.5%	22	25%
Gly	1	2%	1	1%	1	1%
Ile	0	0%	7	8%	2	2%
Lys	1	2%	1	1%	10	11%
Pro	0	0%	4	4.5%	1	1%
Ser	0	0%	7	8%	7	8%
Tyr	0	0%	2	2%	2	2%
Ala	4	8%	3	3%	2	2%
His	11	22%	0	0%	1	1%
Leu	17	34%	2	2%	3	3%
Phe	7	14%	1	1%	1	1%
Val	9	18%	4	4.5%	3	3%
Asp	0	0%	1	1%	3	3%
Cys	0	0%	3	3%	1	1%
Met	0	0%	2	2%	0	0%
Ter	0	0%	6	7%	0	0%
Thr	0	0%	2	2%	0	0%
Trp	0	0%	1	1%	0	0%
Total	50		91		91	

Table 5.10 Codons arising as a result of mutation in the three genes of interest

	β-globin		*PAX6*		*APC*	
	n		n		n	
Stop	0	0%	33	37%	25	28%
Fr.	0	0%	29	32%	65	72%
In frame	0	0%	2	2%	0	0%
Arg	3	6%	5	5.5%	0	0%
Asp	4	8%	2	2%	0	0%
Gln	2	4%	2	2%	0	0%
Glu	2	4%	0	0%	0	0%
Gly	1	2%	0	0%	0	0%
Leu	1	2%	6	7%	0	0%
Met	5	10%	0	0%	0	0%
Pro	16	32%	1	1%	0	0%
Ser	6	12%	4	4.5%	0	0%
Tyr	9	18%	0	0%	0	0%
Val	1	2%	3	3%	0	0%
Ala	0	0%	1	1%	0	0%
His	0	0%	1	1%	0	0%
Ile	0	0%	1	1%	0	0%
Thr	0	0%	1	1%	0	0%
Total	50		91		90	

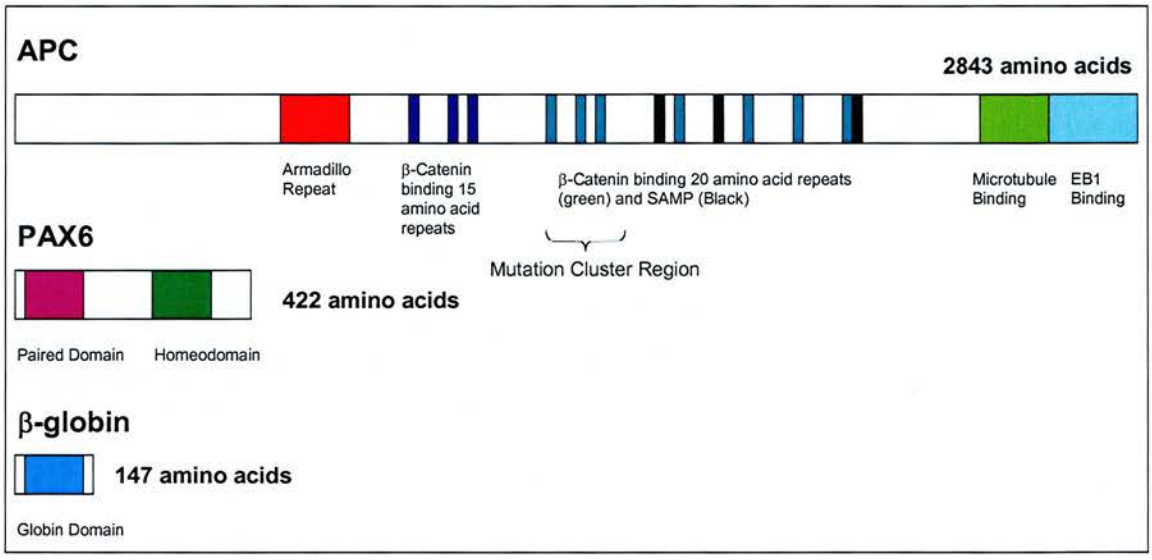

Figure 5.1 Structure of the APC, PAX6 and β-globin proteins with conserved regions indicated

Table 5.11 Distribution of *de novo* mutations in codons and domains

HBB			PAX6			APC		
Codon	No. Mutations		Codon	No. Mutations		Codon	No. Mutations	
25	1	Globin Domain	1	1		170	1	
28	2		17	1		174	1	
29	4		22	1		177	1	
33	3		29	2		181	1	
43	5		35	1		216	2	
59	4		37	2		232	2	
64	8		39	1		243	1	
99	1		40	3		244	1	
67	1		44	1		302	1	
68	3		47	1		452	1	
69	1		53	1		473	1	
72	1		55	1	Paired Domain	482	1	
82	1		56	2		498	1	
88	1		61	1		520	1	
86	1		67	1		554	2	Armadillo Domain
89	2		68	1		585	1	
92	1		73	1		591	1	
93	3		87	1		625	1	
99	5		88	1		650	1	
107	4		95	1		658	1	
116	1		103	1		722	1	
142	1		111	2		753	1	
143	1		113	1		764	1	
			118	1		844	1	
			119	1		868	1	
			123	1		876	3	
			135	2		932	1	
			143	1		935	1	
			150	1		974	1	
			152	1		1000	1	
			156	1		1035	1	15aa rpt
			167	1		1055	1	
			180	1		1061	9	
			181	1		1062	1	
			199	1		1063	1	
			203	4		1066	1	
			205	1		1067	2	
			230	1		1114	1	
			237	2		1125	2	
			240	7	Homeodomain	1170	1	15aa rpt
			247	1		1192	1	
			255	1		1193	1	
			258	1		1259	1	
			261	4		**1309**	**17**	Mutation Cluster Region
			267	1		**1324**	**1**	
			273	1		**1366**	**1**	
			280	1		**1367**	**1**	
			283	1		**1378**	**1**	
			292	1		**1392**	**1**	20aa rpt
			297	2		**1396**	**1**	
			301	1		**1397**	**1**	
			317	5		**1398**	**1**	
			355	1		**1450**	**1**	
			363	1		**1463**	**1**	
			378	1		**1464**	**1**	
			381	1		1533	1	
			384	1		1833	1	
			391	1		1919	1	
			413	1		2209	1	
			422	1				
			423	6				

5.3.3 Sequence conservation and polymorphism in *APC*

Due to the high degree of sequence conservation between the human and chimp genomes (>99%), comparisons between gene sequences can be highly informative in identifying biologically important changes that have occurred since the two lineages diverged around 6 million years ago. In this way it is possible to identify genes that have exhibited accelerated amino acid substitution, relative to other genes, since we diverged from our most common ancestor. Clark *et al.* identified accelerated evolution in genes associated with olfaction and nuclear transport using human-chimp-mouse orthologous gene trios (Clark *et al.*, 2003).

If sequence context and genomic location influence the rate of mutations observed within a gene, *APC* may be located in one of these 'faster' mutating regions. Comparison of the human and chimp *APC* sequences has been used to identify mutations that are fixed in each lineage, and are therefore the product of mutation rate and positive selection. To identify changes that are specific to the divergence in each lineage we compared the human-chimp aligned sequence to mouse. We have additionally compared sequence variants that result in colorectal cancer with amino acid substitutions that have accumulated since the divergence of humans from chimps.

The human-chimp aligned cDNA sequences contained exons 1 to 15, excluding the alternatively spliced exons 9 and 14a that were not present in the chimp sequence. A gap between contigs AADA01173489 and AADA01178933 in the chimp sequence at exon 15 missed 63 base pairs of sequence.

In total there are 38 mismatches between human and chimp sequences (Table 5.12), excluding the alternatively spliced exons, and 63 base pair gap in exon 15. Of these, 29 (76%) were base mismatches, and 9 (24%) were insertions/deletions. The 9 insertion/deletions were all complete codons (i.e. 3 base pairs), 2 codons (6bp) were deleted from the chimp sequence (in exon 4), and 1 (3bp) inserted into the human (exon 15).

Table 5.12 Details of mismatches in the Human-Chimp alignment of *APC* (Del - deletion, Ins – insertion)

Exon Position (Human sequence)	Event	AA Change	Lineage
123	Del	In frame del	Chimp
126	Del	In frame del	Chimp
429	C/A	Pro-Gln	Human
545	G/A	Silent	Human
872	T/C	Silent	Human
940	G/A	Silent	Chimp
1171	G/A	Arg-His	Human
1212	T/G	Silent	Chimp
1243	C/A	Pro-Thr	Human
1262	T/C	Silent	Human
1330	C/A	Pro-Thr	?
1569	A/G	Asp-Gly	Chimp
1589	C/T	Arg-Cys	Chimp
1678	G/T	Silent	Human
1756	T/G	Silent	Chimp
1761	C/T	Silent	Human
1822	T/A	Val-Asp	?
1933	T/C	Silent	Human
1960	G/A	Silent	Human
1978-1979	Ins	Asn	Human
2060	A/C	Asn-His	Chimp
2063	C/T	His-Tyr	Chimp
2073	G/C	Gly-Ala	Human
2295	T/A	Ser-Thr	Chimp
2451	C/A	Silent	Chimp
2468	A/G	Silent	?
2603	A/C	Glu-Ala	Chimp
2620	T/A	Phe-Ile	Human
2690	C/G	Ala-Gly	Human
2749	A/G	Silent	Chimp
2782	T/A	Silent	Human
2786	C/T	Silent	Chimp

Two deleted codons (6bp) and 2 base pair mismatches were present at the 5' end of the gene (exons 1-14) (21% of mismatches), the remaining 27 mismatches, and single inserted codon (3bp) were seen in exon 15 (79%). Within the mutation cluster region, only a single mismatch change was observed, at codon 1330; A CCT (Proline) in human, and an ACT (Threonine) in chimp. The lineage of this change could not be determined using the mouse sequence, as codon 1330 is a GCT (Alanine) in the mouse *APC* gene. At this same position an 8bp deletion, and 1bp insertion have been recorded in an FAP and a hepatoblastoma patient respectively, highlighting the importance of this region in the APC protein. Threonine is a hydrophilic amino acid and forms hydrogen bonds, whereas Proline is aliphatic and hydrophobic. This change (Proline to Threonine) is predicted by Polyphen (http://www.bork.embl-heidelberg.de/PolyPhen/) to be 'potentially damaging'. However, alignments used by this program show that the rat and xenopus APC proteins (P70478 and P70039) also have a Threonine at this position. This suggests that it is the human sequence that has acquired the change.

From 29 mismatched base pairs (excludes ins/dels), 15 (52%) resulted in a synonymous/silent change, and 14 (48%) in a nonsynonymous change. Only 2 of the 14 (14%) nonsynonymous changes were present in the *APC* database as human mutations, and are therefore probable polymorphisms. At codon 1171, the exact same change (arginine to histidine, G-A transition at a CpG dinucleotide) was seen in a patient with gastric cancer. At codon 1822 a Valine to Aspartate change is observed between chimp and human. Its lineage is unknown. The *APC* database records three cases of an Aspartate to Valine change in FAP patients at this position (D1822V), which is believed to be a polymorphism. Although it is often assumed that a change observed between human and chimp is unlikely to cause disease, as selection has not caused its removal, in this case the Valine allele is believed to have a protective effect against CRC, especially when combined with a low-fat diet (Slattery *et al.*, 2001b).

In total 8 synonymous changes were observed in the human lineage, and 6 synonymous changes in the chimp lineage. 6 non-synonymous changes were

observed in the human lineage, and 6 non-synonymous changes in the chimp lineage. The remaining 1 synonymous and 2 non-synonymous changes were of unknown lineage. K_A/K_S (number of non-synonymous differences per non-synonymous site/number of synonymous differences per synonymous site) is a measure of selection acting upon a sequence. $K_A/K_S < 1$ is indicative of negative/purifying selection, >1 indicates positive selection, and $\simeq 1$ indicates neutral evolution. As K_A/K_S is used as a measure of the rate of protein evolution, after controlling for mutation rate (Hurst, 2002), it is a useful indicator of selection pressures. Usually, $K_A << K_S$ and therefore $K_A/K_S < 1$, however in a few instances (i.e. immune system genes) $K_A/K_S > 1$, this is strong evidence for positive selection. In the human lineage K_A/K_S for *APC* is 0.285, and in the chimp lineage 0.378 (data for number of synonymous and non-synonymous sites in *APC* is taken from (Bustamante *et al.*, 2005). These figures indicate that in both lineages the *APC* gene is subject to purifying selection. Since the value of K_A/K_S is lower in the human lineage than in the chimp, there is no evidence of accelerated amino acid substitution in the *APC* gene since the divergence of humans from our most common ancestor.

Data on coding single nucleotide polymorphisms in *APC* was analysed to give a figure of P_N/P_S (number of non-synonymous polymorphisms per non-synonymous site/number of synonymous polymorphisms per synonymous site). Using the randomly found polymorphisms from Bustamante *et al.* P_N/P_S is 0.378. By comparing P_N/P_S and K_A/K_S, inferences can be made about selection acting upon a sequence, i.e. molecular adaptation in evolutionarily conserved genes, and differentiation between weak and strong purifying selection. Genome-wide analysis has shown that the ratio of non-synonymous to synonymous differences is smaller than the ratio of non-synonymous to synonymous polymorphisms (Bustamante *et al.*, 2005). This suggests that a large proportion of amino acid variation in the human genome is slightly to moderately deleterious. From data presented here K_A/K_S (0.285) < P_N/P_S (0.378). There is therefore an excess of amino acid variation relative to divergence. This suggests a weak purifying selection is acting upon mildly deleterious mutations to increase the proportion of non-synonymous polymorphisms relative to the proportion of non-synonymous fixed differences.

We have used the PAML program (default settings) to calculate K_A and K_S for the *APC* gene using the mouse orthologous sequence (Table 5.13). This has allowed us to compare *APC* with a number of other genes analysed in published articles that have calculated K_A and K_S in the same way.

Table 5.13 Values of Ka and Ks for the Human-Mouse alignment of *APC*

	Human-Mouse
Ks	0.5221
Ka	0.0464
Ka/Ks	0.089

Smith and Eyre-Walker (Smith and Eyre-Walker, 2003) have calculated K_A and K_S values for disease and non-disease genes using human-mouse alignments (with PAML default settings). They found that disease genes are evolving faster than non-disease genes at both synonymous and non-synonymous sites. This is expected to be due to non-disease gene evolution being more affected by lethal mutations than disease genes. Disease genes (less affected by lethal mutations) therefore evolve faster because mutations have a discernible affect on phenotype. From human-mouse alignments, average values of K_A/K_S for non-disease genes and disease genes were 0.0913 ± 0.0022 and 0.1135 ± 0.0052 respectively. These values are higher than 0.089 for the *APC* gene. Similarly, Chuang and Li have used K_S values to compare mutation rates in families of genes. Using PAML (default settings) they estimated, from human-mouse alignments that the genome average for K_S is approximately 0.337, significantly lower than the value we have obtained for *APC*. Genes in 'Hot' or rapidly mutating regions of the genome were found to have K_S values of 0.346-0.468, whilst genes in cold regions had K_S values of 0.220-0.326. Thus, *APC* would appear to be within, or an integral part of a rapidly mutating region of the genome.

5.4 Discussion

In this chapter, a variety of *in silico* techniques have been used to investigate further the possibility of instability in the *APC* gene. The consequence of *APC* instability on the prevalence of sporadic FAP has been determined by directly comparing the *de novo* mutation rate of *APC* with a number of other disease genes. To determine if there might be any mutational mechanisms at play, other than sequence instability, that are influencing the frequency of *de novo APC* mutations, the mutations of *APC* have been analysed alongside *de novo* mutations arising in other disease genes. Lastly, the degree of sequence conservation and polymorphism in *APC* has been ascertained, as this provides historical evidence of the selection and mutational influences that *APC* is under.

To establish and compare *de novo* mutation rates for a set of genes, the 'direct method' is the most straightforward. Unlike the 'indirect' method (Vogul and Motulsky, 1986) there is no need to calculate the average fertility of patients in relation to the population average. However, establishing the *de novo* mutation rate for a gene still requires a large amount of epidemiological data. This includes the incidence of cases in the population where parents and other family members are unaffected i.e. sporadic/*de novo* cases. The number of births in the same time period must also be obtained. If a condition cannot be identified at birth, then prevalence data is used. However, if life expectancy is reduced in mutation carriers, then prevalence figures must be adjusted accordingly, to account for the reduction in observed prevalence due to death. Despite the importance and significant interest of determining the mutation rate for a disease gene, as a result of the difficulties involved in obtaining the necessary data, very few genes have been analysed in this way.

One of the main difficulties with determining mutation rates by the direct method is the differences in the number of nucleotides per gene, as was eluded to in 5.3.1. The analysis of the *de novo* mutation rate of the *APC* gene, and its comparison with other disease genes must be done in the context of a number of other factors that may

introduce error. Illegitimacy will introduce errors if there are very few sporadic cases observed amongst a majority of familial ones. In the more recent studies, paternity will have been confirmed, but this is not always the case in earlier estimations of mutation rate. If the penetrance of a condition is not 100%, then not all sporadic cases in a population will be determined. In this comparison all conditions are believed to have complete, or virtually complete penetrance. Different autosomal dominant conditions caused by mutations in the same gene will cause an elevated mutation rate to be recorded if both conditions are considered together. For this reason only cases of unstable Hb disease and Apert syndrome were used to calculate the mutation rate of *β-globin* and *FGFR2* respectively, despite the fact that mutations in *β-globin* and *FGFR2* can also cause other conditions. When comparing differences in mutation rates between genes that lead to disease, selection against a condition must be strong enough to account for the fraction of sporadic cases among all cases. If a disease phenotype is so severe that it causes virtually no transmission of the mutation to the next generation, then almost all cases will arise due to *de novo* mutations (this is true for Prader-Willi syndrome). In this situation a high *de novo* mutation rate must exist if a condition is to persist (as is seen in Table 5.2). In less severe diseases, such as FAP, Apert syndrome, Neurofibromatosis, VHL disease, HbM and unstable Hb disease, a variably large fraction of cases will be new mutants, and it will be roughly proportional to the selective disadvantage of the trait. If the trait has no bearing on reproductive fitness, then only a very low level of *de novo* mutations will be required.

In addition to the caveats listed above, it should be noted that mutation rates determined by observing a phenotype, i.e. the presence or otherwise of a disease, cannot reveal the true mutability of a gene, because they only consider mutations that produce a visible phenotype. Silent mutations and embryonic lethal mutations are not considered. However, mutation rates calculated in this way can provide an indication of any differences in mutation probability that may exist between genes, especially when they have been corrected for differences in gene size.

After correcting for gene size, three genes had mutation rates greater than *APC*. The *VHL* gene had a rate 7 fold greater than *APC*. The study in which the mutation rate for *VHL* was calculated was carried out in East Anglia (Maher *et al.*, 1991). A relatively small population size from which cases were determined (485,437 compared with over 2 million in the FAP study) may have biased the study to give an inaccurately high mutation rate. In a small study population the chances of a single large kindred being responsible for a large number of cases is much more likely. Additionally, this study was carried out prior to the *VHL* gene being identified. Phenotype was used to infer gene mutation, but unlike the FAP study, true mutation carriage could not be confirmed. Therefore, cases cannot all be definitely attributed to mutations in *VHL*. The *FGFR2* gene also had a mutation rate greater than *APC*. Aperts syndrome is a severe phenotype, and few affected individuals have children. It appears that the disease is only able to persist because of the high susceptibility of just two adjacent nucleotides (Cohen, Jr. *et al.*, 1992) (including one at a CpG dinucleotide) to mutation in the male germline (Moloney *et al.*, 1996). The mutation rate considered here is therefore only representative of a very small proportion of the *FGFR2* gene. The mutated nucleotides are involved in the interaction between FGFR2 and its ligand (FGF2), but the possible effect of sequence context on mutation rate has not been determined.

Of the genes examined here, the *NF1* gene is most comparable with *APC*. *NF1* is a similar sized gene to *APC*, is also a tumour suppressor gene, and estimates of the effect on reproductive fitness of being a mutation carrier is around 0.8 for both conditions (Samuelsson and Akesson, 1988;Bisgaard *et al.*, 1994). The mutation rate of *NF1* from this study is 4 fold greater than in *APC*, although earlier reports of *NF1* mutation rate have given the highest known rate for a human disorder (1×10^{-4}) (Crowe FW *et al.*, 1956).

From these data alone it is not possible to conclude that the *de novo* mutation rate of the *APC* gene is significantly higher than for other genes where relevant data is available. However, the rate of occurrence of new mutations in *APC* is comparable

with that of other similar genes for which high new mutation rates have been previously proposed.

In a Dutch cohort of sporadic colorectal carcinomas, 61% of mutations were found to be missense point mutations, and multiple mutations were seen in 55% of cases (52% of individuals with a truncating mutation also had a missense change) (Miyaki *et al.*, 1995). This study suggests that missense mutations are under-reported in FAP, leading to an ascertainment bias. Their relevance in the context of *APC* is not well understood. As a result, it has not been possible to accurately determine what proportion of *de novo* mutations are missense changes, and what influence, if any, these have on the *de novo* mutation frequency of *APC*. A lack of controls and limited study populations can also hinder comparisons between databases.

It is unsurprising that there are no observable differences between *de novo* and familial mutations in *APC,* if one considers that all familial mutations were sporadic mutations when they occurred for the first time. Likewise, sporadic mutations will become 'familial' if they are inherited. As discussed above, each generation *de novo* mutations will replace those familial mutations that have been lost from the gene pool due to a reduction in reproductive fitness. These *de novo* mutations are subsequently passed on to further generations creating new FAP families.

Detailed analysis of the mutations within a gene, with special attention to any biases in the data can provide information on mutational hotspots, and mechanisms by which mutations arise. It has been proposed that new germline mutations in *APC* are not premeiotic events (because no parental bias is observed in mutation origin), but instead occur during meiotic division (Ripa *et al.*, 2002). This is an unusual finding, as most other genes studied are not susceptible to genetic changes at meiosis, with mutations instead arising during the mitotic divisions preceding meiosis, and therefore showing parental origin bias. The error rate of mutations occurring during DNA synthesis is known to be dependent on the DNA sequence being synthesised, and the polymerase used. Polymerase-δ has been proposed to play a role in semiconservative DNA synthesis and repair during the early stages of meiosis (Plug

et al., 1997), whilst polymerase-ε is believed to replicate DNA at mitosis. Ripa *et al.* propose that errors arising during polymerase-δ DNA synthesis might offer an explanation as to why the locations of somatic and germline mutations differ significantly within *APC* (Miyaki *et al.*, 1995;Ripa *et al.*, 2002). It may also provide a mechanism for the emergence of *de novo APC* mutations, if the sequence context, or potentially the genomic region of *APC* is susceptible to mutation by polymerase-δ during oogenesis or spermatogenesis.

By comparing the types and locations of *de novo* mutations in the *APC, PAX6* and *β-globin* genes, it has been shown that the differences that exist between the three loci reflect differences in nucleotide sequence, the biology of the conditions caused by mutations, and the size and structure of the genes. Taking into consideration all of these factors, it has not been possible to identify any likely underlying mechanism that might account for the *de novo* mutations arising in *APC*. Although, polymerase-β induced errors cannot be ruled out. These comparisons are however an important tool for further work into understanding what influences *de novo* mutations. Research into the genetic basis of *de novo* FAP may shed light on the genetic basis of colorectal cancer in general.

Comparisons of the human and chimp *APC* sequences reveal that evolutionary pressures on the two lineages since they diverged are likely to have been similar and purifying. In contrast to this, the value of K_S that has been calculated from the divergence between human and mouse indicate that *APC* is mutating at a rate that is faster than genes in previously designated 'hot' regions of the genome. Hot (genes under positive selection/high neutral mutation rate) and cold (purifying selection/low neutral mutation rate) regions of the genome have been proposed by Chuang and Li to explain the observation of differing neutral mutation rates (Chuang and Li, 2004). They hypothesise that the observed non-random distribution of gene families offers the possibility that selection is acting at the level of gene location in the genome. Neutral mutation rates are often used to determine how fast a sequence is evolving, since they are not subject to the usual selection pressures present at non-neutral sites. The Mouse Genome Sequencing Consortium (Waterston *et al.*, 2002) used

alignments of the mouse and human genomes to study neutral mutation rates, reporting variation in the substitution rate across the human genome, indicating 'substantial regional fluctuation'.

When APC protein evolution is measured, controlling for this high neutral mutation rate, the resulting value is lower than previously reported for disease genes (Smith and Eyre-Walker, 2003b), suggesting a striking increase in selection against non-synonymous substitutions. Hurst and Smith (Hurst and Smith, 1999) show that essential genes have significantly lower K_A/K_S ratios than non-essential genes. Polymorphism data suggests that purifying selection is indeed acting upon *APC* and increasing the proportion of non-synonymous polymorphisms relative to the proportion of non-synonymous fixed differences. These data suggest that *APC* may be a highly mutable gene, or lies in a highly mutable region of the genome, but that purifying selection is acting as a constraint mechanism maintaining the DNA sequence 'constant'. This increase in mutation rate may stem from local sequence complexity within the *APC* gene, or may arise as a result of structural and conformational features of the DNA in this region of the genome. Differences in structure and conformation of the DNA may lead to differential mutation rates by providing different levels of exposure to mutagenic and DNA damaging agents, and enzymes involved in DNA processing. A possible explanation as to why this affect has not been detected in the human-chimp comparison is that there is not sufficient evolutionary time between the two species, unlike the human-mouse comparison.

Although the human and chimp *APC* genes appear to have been subject to similar evolutionary pressures, their alignment has highlighted some interesting differences. Using the context of human-chimp divergence to analyse two amino acid changes from our alignment that are also present in the *APC* mutation database, suggests that both are likely to be human polymorphisms, and not causative in the conditions (indeed, D1822V is predicted to have a protective effect (Slattery *et al.*, 2001b)). Amino acid substitutions associated with human disease tend to occur at sites that are more highly conserved in other organisms (Smith and Eyre-Walker, 2003c), and so they can give novel insights into the selective relevance of genetic changes.

Alternatively, as has been proposed by Gao and Zhang, a compensatory mutation hypothesis might explain the fixation of disease causing mutations in other organisms (Gao and Zhang, 2003). Within the mutation cluster region, only a single mismatch change was observed, indicating high conservation at this region, likely due to the consequences of mutations that occur here.

Data presented in this chapter provides some *in silico* evidence to support the hypothesis of instability in the *APC* gene sequence that has been investigated in Chapter 3. By comparing the value of K_S that we have calculated for *APC* with figures previously calculated across the genome, we can conclude that the mutation rate at neutral sites is indeed unusually high. In contrast, polymorphism and K_A values indicate that *APC* is subject to purifying selection pressures that eliminate mutations. If the majority of mutations arising in *APC* are only mildly deleterious (i.e. missense changes), they may affect one specific role of APC, and will be selected against. On the other hand truncating mutations in *APC* (which can be assumed to arise occasionally as a consequence of a high mutation rate) may offer a selective advantage to the cell by disrupting vital roles such as Wnt signalling, and are therefore likely to be subject to positive selection in the short-term/at the cellular level.

Possession of a germline truncating mutation in *APC* is likely to result in a reduction in reproductive fitness. For this reason the gene must be generating a significant number of *de novo* mutations each generation. Our data on the *de novo* mutation rate of *APC* is not conclusive, due to the rarity of suitable data with which to make comparisons. However, we believe our evidence indicates that there is a high mutation rate for *APC* in the germline as well as somatic cells. This is supported by the lack of specific mutation types amongst *de novo* mutations compared with familial mutations, or sequence specificity that might have implicated a particular process in the generation of new germline *APC* mutations.

Chapter 6

Measurement of DNA Base Excision Repair in Live Cells

6.1 Introduction

It has recently been shown that loss of DNA base excision repair function leads to the development of colorectal cancer (Al Tassan *et al.*, 2002;Jones *et al.*, 2002;Lipton *et al.*, 2003;Cheadle and Sampson, 2003;Farrington *et al.*, 2005). Current methods for identifying patients harbouring deficient BER involves the identification of mutations in the genes known to be involved, such as *MUTYH* and *OGG1* (Sampson *et al.*, 2003;Vandrovcova *et al.*, 2004;Russell *et al.*, 2005). Whilst these are reliable methods, they provide no information about a mutation's functional consequence.

Currently, assays for BER activity are technically difficult, time consuming and expensive. A number of techniques use cell extracts rather than live cells, these include the '*In vitro* Cleavage Assay' where whole cell extracts are incubated with an antibody of interest (Parker *et al.*, 2002;Alhopuro *et al.*, 2005), and the highly complex '2-Photon Fluorescence Cross-Correlation Spectroscopy' where uracil DNA-glycosylase activity is measured in cell protein extracts (Collini *et al.*, 2005). An example of an assay that utilises live cells is the 'Host Cell Reactivation Assay' which is based on the expression of a repair gene and/or repair gene mutants of interest, alongside a reporter gene (Carreau *et al.*, 1995). The reporter gene is treated with a DNA damaging agent *in vitro*, and then introduced into cells by transfection. Resulting gene expression thus reflects the repair ability of the cells. Whilst this assay gives quantitative information about the activity of the repair proteins of interest *in vivo*, complicated vectors need to be constructed. Clearly, there is an increasing need for simple and reliable approaches to compare and monitor BER

capacities in various types of cells, under various physiological and pathological conditions.

At this time there are reports of three functional assays that utilise the green fluorescent protein (GFP) to determine DNA repair function. In 1999 Pierce *et al.* developed a recombination reporter system to assay chromosomal gene conversion events in response to double strand breaks (DSBs) (Pierce *et al.*, 1999). A vector containing the GFP gene with an endonuclease site inserted into it, alongside a gene coding for the corresponding endonuclease was transfected into cells of interest. When expressed, the endonuclease cleaved the GFP protein, however, functional homologous repair using a linked donor GFP fragment could restore GFP expression and this was quantified by flow cytometry two days after transfection.

BER has been further described by two sub-pathways; 'short patch' BER (SP-BER) and 'long patch' BER (LP-BER). In SP-BER only a single nucleotide is excised and replaced, but in LP-BER at least two nucleotides are excised and replaced by gap-filling (Klungland and Lindahl, 1997). Sattler *et al.* set out to quantify the proportion of BER activity that is specifically attributable to LP-BER (Sattler *et al.*, 2003). They constructed a plasmid that contained a single 8-Oxoguanine residue within an enhanced GFP (EGFP) gene, close to an inserted stop codon. If 8-OxoG was corrected, LP-BER would remove the stop codon, and restore EGFP expression, but SP-BER would not. Cells were incubated for 18 hours following transfection with the plasmid, and fluorescence was then quantified by flow cytometry.

A novel method has been successfully developed by Lei *et al.* to examine mismatch repair activity in a variety of cells with and without known MMR defects (Lei *et al.*, 2004). A mismatch that knocked out gene expression was introduced into the start codon of the EGFP gene. Following transfection of cells with this mutant vector, levels of EGFP protein expression were detected by flow cytometry. The number of cells repairing the start codon mismatch could then be calculated for each cell line, giving a measure of mismatch repair activity.

Data presented in this chapter aims to develop an assay that would be useful for comparing and monitoring total BER activity in a number of live cells, under a number of different conditions. To our knowledge there is no assay available to directly measure total BER activity in live cells at this time. Our approach has been to make use of a vector carrying the EGFP gene, and use flow cytometry for the high-throughput analysis of a large number of cells, in a similar fashion as Lei *et al.*, 2004.

In the development of this assay we were able to take advantage of two primary cell lines, the colorectal cancer cell line, VACO425 (Parker *et al.*, 2002) and the fibroblast cell line, 2630 (See Chapter 3) that are known to be BER null. We also analyse other MMR null and control colorectal cancer and fibroblast cell lines that are available to us (Table 2.1).

6.2 Methodological Overview

6.2.1 Biological Material

Colorectal cancer cell lines VACO425 (BER null, see (Parker *et al.*, 2002) and Chapter 3), HCT116 (MMR null, see (Papadopoulos *et al.*, 1994) and Chapter 3) and SW480 (BER and MMR proficient, (Leibovitz *et al.*, 1976)), and fibroblast cell lines 2630 (BER null), 825 (MSH2 heterozygote) and 439 (BER and MMR proficient) (see Chapter 3 and 2.1.1) were maintained as outlined in 2.1.2.

6.2.2 Site Directed Mutagenesis of pEGFP-C1 vector

The 4731 base pair pEGFP-C1 vector (Becton Dickinson Biosciences) encodes an enhanced GFP (EGFP) gene (nt 613 – nt 1410) under the control of a human CMV promoter, see Figure 6.1. Site directed mutagenesis (2.6) was used to introduce a stop codon into the 5' end of the GFP gene. Codon 45 of the EGFP gene (vector position nt 745 – nt 747) was chosen, as mutation of the wild-type sequence, CTG (Leucine), to a TGA stop codon, was predicted to disrupt EGFP expression, and gave rise to a G residue upstream of an AA dinucleotide, the most likely sequence to be mutated in the presence of deficient BER (Al Tassan *et al.*, 2002) (Figure 6.2, top panel). Consequently, a G:C-T:A transversion mutation at the codon 45 TGA, stop, would give rise to a TTA (Leucine) codon, therefore restoring the original amino-acid at this position, ensuring wild-type EGFP expression.

Primers used to mutate codon 45 are detailed in 2.6.1. Following mutagenesis of codon 45, 12 colonies were picked using sterile pipette tips. Colonies were stabbed into a 96 well stock plate containing 200µl L-Broth containing kanamycin (30µg/ml), and grown with shaking for 8 hours prior to being stored at 4°C. Stock cultures were miniprepped and sequenced as detailed in 2.2.2 using the primers listed below, to ensure that the mutation had been successfully inserted, and that no other

mutations had been introduced to other parts of the vector during the site directed mutagenesis process:

EGFP-N: CGTCGCCGTCCAGCTCGACCAG (nt 679 to nt 658, to check sequence upstream of EGFP gene)
EGFP-MUT: GGTGCGCTCCTGGACGTAGCC (nt 906 to nt 886, to ensure mutation had been successfully inserted)
EGFP-CF: CATGGTCCTGCTGGAGTTCGTG (nt 1266 to nt 1287, to check sequence downstream of EGFP gene)
EGFP-CR: CACGAACTCCAGCAGGACCATG (nt 1287 to nt 1266, to check whole EGFP gene, including mutation at codon 45)

Sequencing from these primers provided coverage from nt 4581 to nt 1990, incorporating approximately 50% of the vector, and the entire EGFP gene. Of the 12 colonies sequenced, 9 had the appropriate TGA sequence at codon 45, with no other mutations introduced in other areas of the vector. Of these 9, two were picked from the stock plate and grown in a 5ml overnight culture of L-Broth containing kanamycin (30μg/ml). This was pelleted at 3000rpm for 5min in a Sorvell RT6000 centrifuge, and resuspended in 0.85ml L-Broth + 0.15ml glycerol to make a glycerol stock for long-term storage at –80°C.

Glycerol stocks were used to streak agar plates containing Kanamycin, and incubated overnight at 37°C. A single colony was picked, and plasmid DNA isolated for transfection by maxiprep (Qiagen).

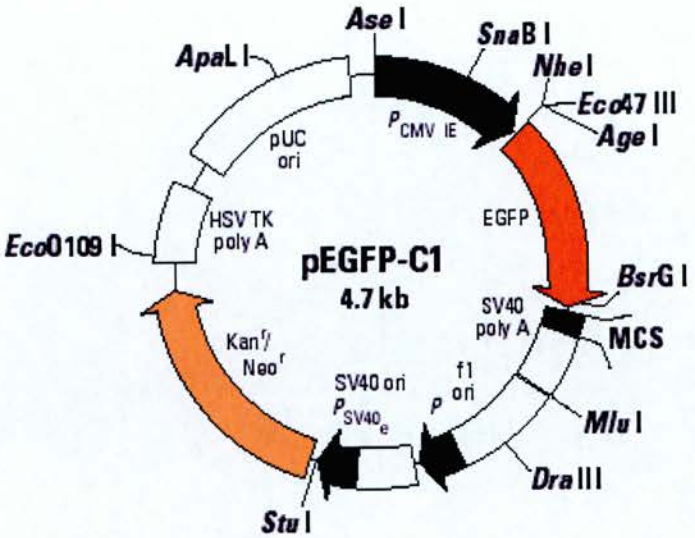

Figure 6.1 pEGFP-C1 vector (Becton Dickinson Biosciences).

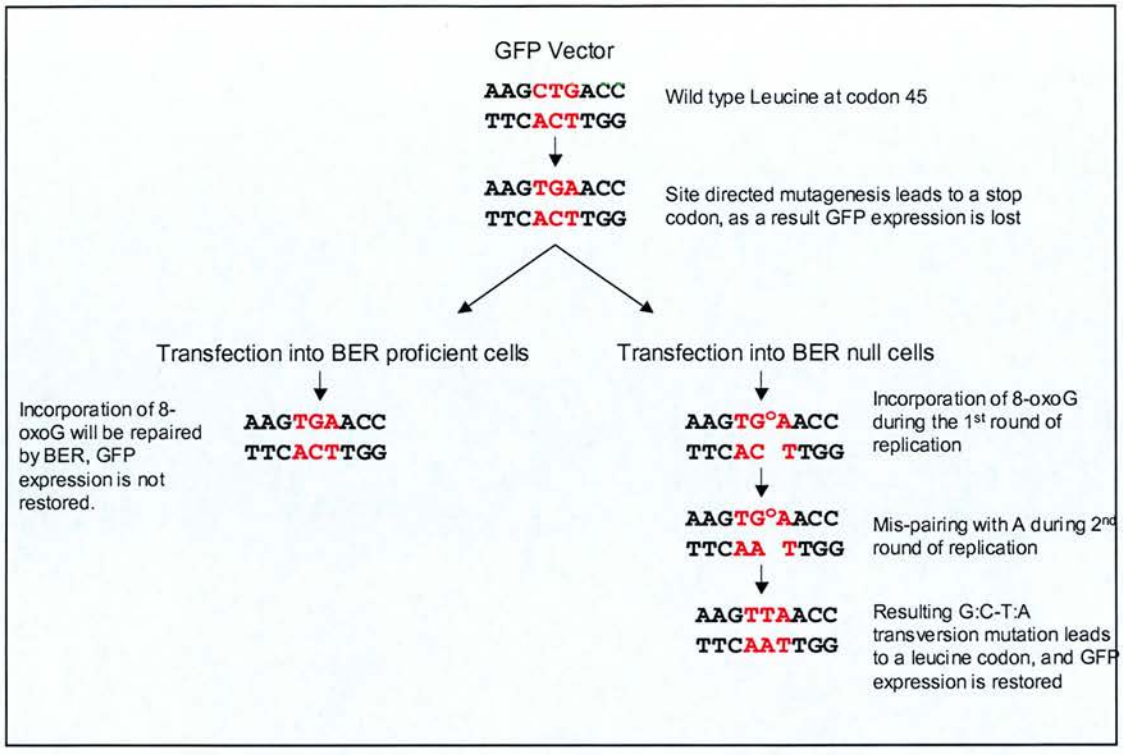

Figure 6.2 Schematic diagram of assay to detect BER in live cells. Top panel shows the codon change resulting from site directed mutagenesis of the pEGFP-C1 vector. Bottom panel represents the proposed mechanism by which BER null cells might revert the EGFP mutation, so that expression is restored.

6.2.3 Transfection of cell lines

On day 1 cells were plated at a density of ~ 2 x 10^5 in 2 ml media, into 6 well dishes for transfection (2.5.3). In three parallel experiments 6 well dishes were set up with 3 wells of cells for transfection with mutant plasmid DNA (6.2.2), unmutated (wild-type) plasmid DNA, and an untransfected control, as described in 2.5.3. Cells transfected with the mutant and wild-type vectors were co-transfected with 0.6μl (2.745μg/μl) of the pSV-β-Galactosidase control vector (Promega) (2.5.4). Transfection took place on day 2, followed with a media change on day 3.

6.2.4. FACs analysis of transfected cells

Cells were harvested for FACs analysis on day 5, as outlined in 2.5.5. 50μl aliquots of cells re-suspended in PBS were then removed, to control for transfection efficiency by β-Galactosidase enzyme assay system (2.5.4).

A FACS Caliber cytometer, was used to analyse fluorescence levels in all samples, as detailed in 2.5.5. FL1 and SSC parameters were adjusted using the untransfected samples to insure that differences in the amount of auto-fluorescence did not influence results. Wild-type and mutant transfected cells could then be analysed, with reference to the un-transfected sample.

6.2.5 Data Analysis

Fluorescence data was analysed for all cells with $>10^2$ units of fluorescence, as this high level of fluorescence was never seen in un-transfected cells, therefore any transfected cells with this level of fluorescence are assumed to be expressing EGFP. Total EGFP expression was calculated as a product of N (number of cells with fluorescence $>10^2$ units) and I (mean intensity of fluorescence in these cells), for all samples. Total EGFP expression was then corrected for transfection efficiency using levels of β-Galactosidase enzyme activity. Lastly, the mean levels of fluorescence

and standard errors were calculated from the three experiments carried out in parallel (total of 15,000 cells).

Relative EGFP expression was also calculated to determine the proportion of wild-type expression seen in cells transfected with the mutant plasmid. This was calculated as (corrected N_M x I_M)/(corrected N_{WT} x I_{WT}), where M represents cells transfected with mutant vector, and WT represents cells transfected with wild-type vector. Again, the means and standard errors were calculated.

6.3 Results

6.3.1 Optimisation of the method

The transfection procedure was optimised using the wild-type vector on all 3 fibroblast cell lines (825, 2630 and 439) and the colorectal cancer cell line VACO425. It was found that the Lipofectamine™ 2000 (Invitrogen) reagent gave the highest level of transfection efficiency in all cell types (used according to manufacturers instructions), when compared directly with the Lipofectamine™ (Invitrogen) reagent. Cell density was also optimised, with plating 2 x 10^5 cells in 2 mls media, giving approximately 50-60% confluency of cells on day 2 when transfection took place, and the best transfection efficiency, with approximately 50% of cells expressing EGFP at some level (determined by fluorescent microscopy). Higher confluency resulted in a high level of cell death following transfection. As per manufacturers instructions, 4μg vector was found to give the best transfection efficiency, in the 6 well dish format. A media change was introduced on day 3 to feed cells, this had no effect on transfection efficiency, but reduced cell death following transfection. Cells transfected on day 2 were analysed for fluorescence on day 5 by flow cytometry, this 72 hour incubation was included so that the vector could be replicated, and therefore subjected to DNA repair mechanisms and/or deficiencies.

It was hypothesised that BER deficient cells would show higher levels of EGFP expression from mutant constructs than MMR deficient or BER/MMR proficient cells. The initial optimisation included H_2O_2 treatment to induce oxidative damage, and therefore detect higher levels of EGFP expression from mutant constructs (Slupphaug *et al.*, 2003). However, during the optimisation of the protocol, transfected and untransfected cells were treated with 100μM, 200μM and 400μM H_2O_2 for 24 hours as part of the media change step on day three, but the treatment was highly toxic to our cells, and resulted in widespread cell death, so was not pursued further.

During the initial optimisation experiments, detectable fluorescence from the mutant vector was only detected in the BER null VACO425 colorectal cancer cell line. The other cell lines studied were all of fibroblast origin (including the BER null 2630 line), and it was postulated that this effect might be cell type specific. For this reason, 2 additional colorectal cancer cell lines were added for the final set of experiments for comparison. These were, SW480 (MMR and BER proficient) and HCT116 (MMR defective).

6.3.2 β-Galactosidase enzyme activity reveals higher levels of transfection efficiency in colorectal cancer cell lines compared with fibroblast cell lines

To determine the levels of transfection efficiency that could be expected in our six cell lines of interest, they were co-transfected with the wild-type pEGFP-C1 vector and the pSV-β-Galactosidase control vector. β-Galactosidase enzyme activity was then used to correct for differences in transfection efficiencies between the cell lines (2.5.4). The colorectal cancer cell lines VACO425, SW480 and HCT116 showed higher levels of transfection efficiency by β-Galactosidase enzyme assay than the fibroblast cell lines 439, 825 and 2630 (Table 6.1). The efficiency of DNA uptake by transfection varied enormously between cell lines, but the cells of fibroblast origin were all extremely low. In this case it is most likely that differences in proliferation rate can explain this phenomenon, as the tumour-derived cell lines can be seen to be multiplying at much higher rates than the primary fibroblast cell lines. Liposomes are incorporated into cells mainly by endocytosis (Zhou and Huang, 1994;Zabner *et al.*, 1995). High cell proliferation activity correlates well with cellular endocytotic activity, therefore transfection of cationic liposomes, such as used in the lipofectamine 2000 reagent, will be more efficient in highly proliferating cells. This association has been confirmed via β-Galactosidase enzyme assay in a study of oral tumour cell lines (Nakase *et al.*, 2005).

Table 6.1 β-galactosidase enzyme activity in six cell lines. The colorectal cancer cell lines, VACO425, SW480 and HCT116 have increased transfection efficiencies over the other cell lines. Units are absorbance at 420nm.

Cell Line	Co-transfected with Wild-Type vector	Co-transfected with Mutant vector
VACO425	0.027	0.029
SW480	0.073	0.048
HCT116	0.283	0.357
439	0.004	0.006
825	0.004	0.006
2630	0.008	0.002

Despite correcting for transfection efficiency, Figure 6.3 shows that average levels of wild-type EGFP expression in the HCT116 cell line, 72 hours after transfection, were greatly increased compared with the other five cell lines (around 10-fold), including the other two colorectal cancer cell lines. However, when disregarding the HCT116 cell line, β-Galactoidase enzyme activity did effectively control for differences in transfection rates in the other 5 cell lines. The failure of β-Galactosidase enzyme activity to control for the transfection efficiency in HCT116 can be attributed to the threshold of the β-Galactosidase enzyme assay being exceeded by high levels of expression. Figure 6.4 shows a standard curve for β-Galactosidase enzyme activity that was calculated subsequent to our EGFP assay. It shows clearly that whilst absorbance at 420nm increases linearly up to 5 milliunits of β-Galactosidase activity (x axis), after this the curve levels off, and absorbance becomes a less accurate method of determining enzyme activity.

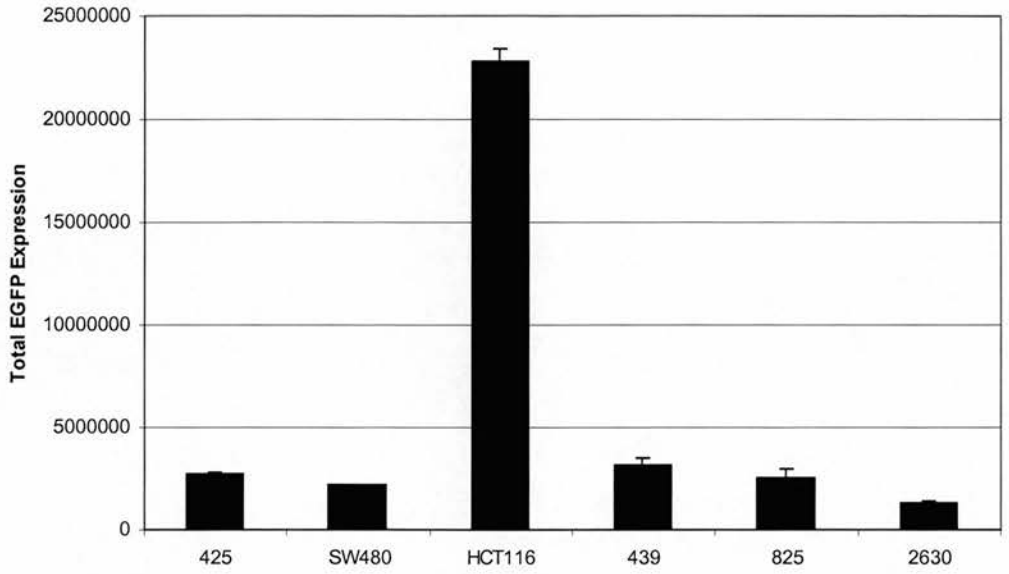

Figure 6.3 Average total EGFP expression levels in 6 cell lines transfected with the wild-type pEGFP-C1 vector. Colorectal cancer cell line HCT116 has around a 10 fold higher level of total fluorescence than the other 5 cell lines. Error bars represent standard errors for three separate experiments of 5000 cells each (total 15000 cells per cell line). Levels of total EGFP expression are arbitrary units

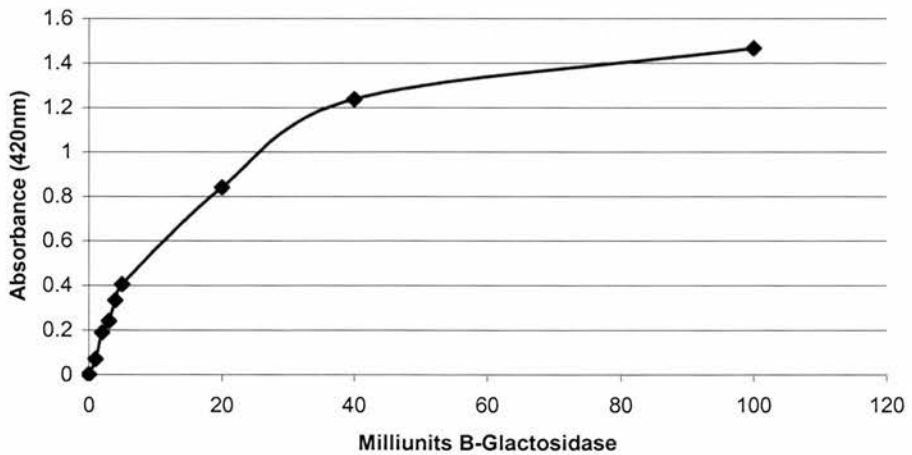

Figure 6.4 Standard curve for β-Galactosidase enzyme activity. Standards of 0, 1, 2, 3, 4, 5, 20, 40 and 100 x 10^{-3} units of β-Galactosidase were used.

6.3.3 VACO425 cell line shows significantly higher mutant E-GFP expression than 5 other cell lines

Total fluorescence was calculated as outlined in 6.2.5, as a product of N (number of cells with fluorescence $>10^2$ units) and I (mean intensity of fluorescence in these cells), this was corrected for transfection efficiency and standard errors from three experiments were calculated. Figure 7.5 shows typical FACs outputs for all six cell types when untransfected, transfected with the wild-type pEGFP-C1 vector, and transfected with the mutant pEGFP-C1 vector.

Figure 6.6 shows the average total fluorescence (level of EGFP expression) from the mutant pEGFP-C1 vector for each cell line. Despite extremely high levels of EGFP expression from the wild-type vector in the HCT116 cell line (6.3.2), no expression was detected from the mutant vector. This provides good evidence that expression of the EGFP gene, from the mutant pEGFP-C1 vector is not detected in the absence of a BER defect, even when it is present in a high proportion of cells, and is not effectively corrected by β-Galactosidase enzyme activity. In the colorectal cancer cell line, VACO425, the average level of fluorescence detected from the mutant vector was 2707 $\pm$ 793 arbitrary units. This is 2.5 fold greater than the level of fluorescence detected in the SW480 CRC cell line (943 $\pm$ 943 arbitrary units), however since standard errors pass through zero for SW480, this may not be a real effect. The level of fluorescence detected in the VACO425 cell line was more than 2500 fold greater than in the HCT116 cell line, or the three fibroblast cell lines, where no fluorescence was detected from the mutant vector. This included the 2630 cell line, known to be defective for BER due to bi-allelically inactivated *MYH*.

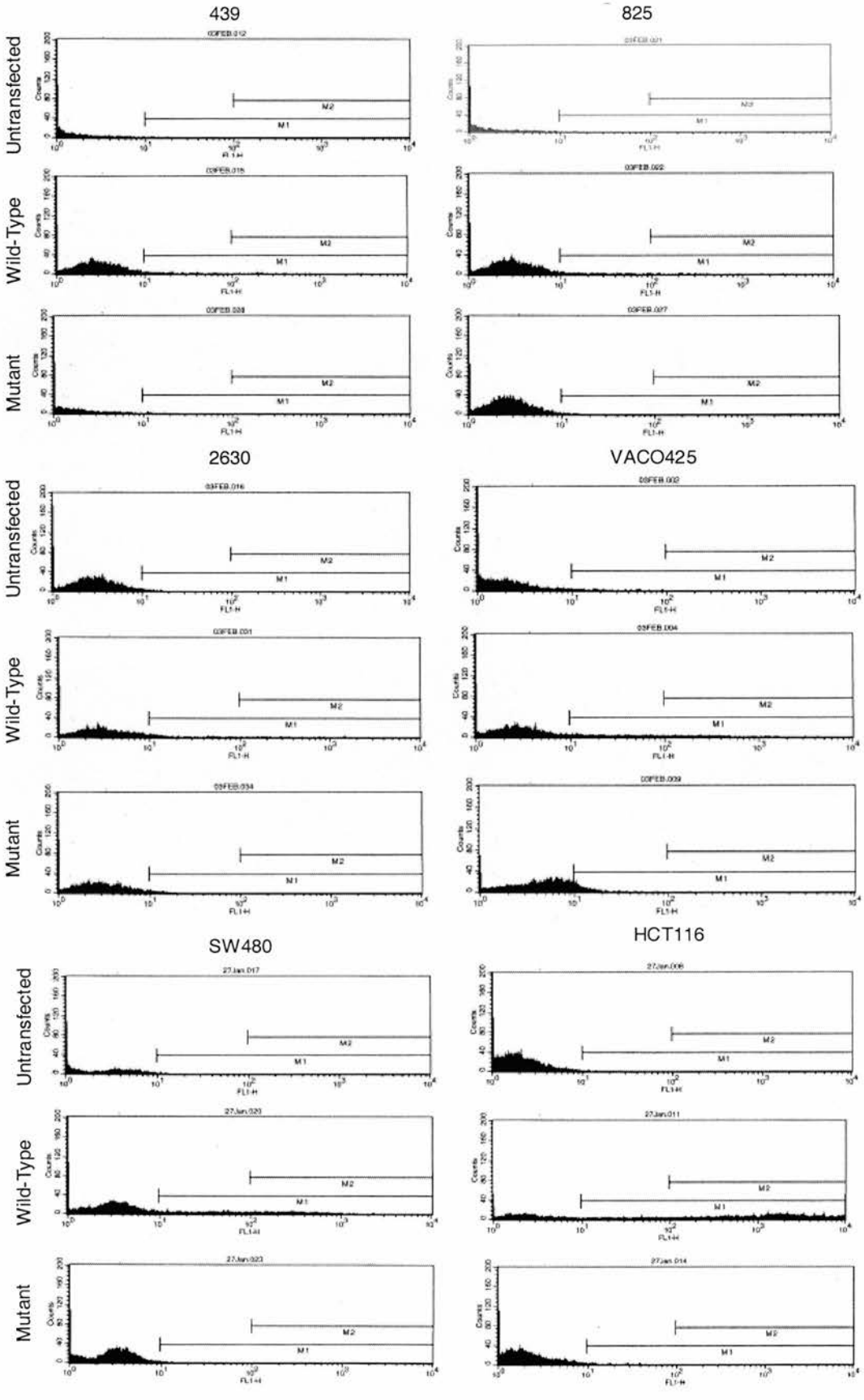

Figure 6.5 Typical FACs data set for each cell line untransfected (top panel), transfected with the wild-type vector (middle panel) and transfected with mutant vector (bottom panel)

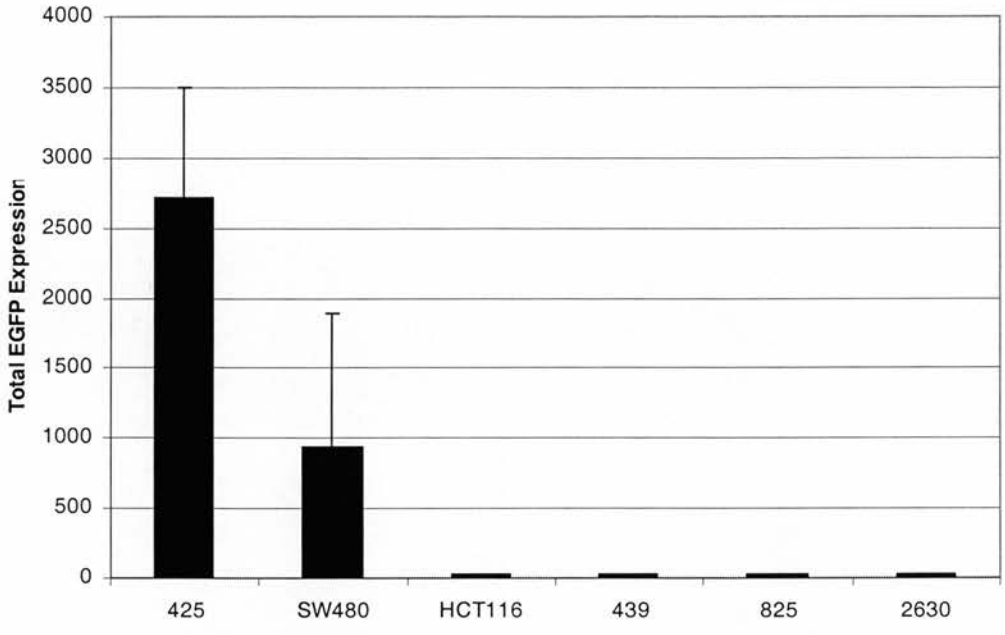

Figure 6.6 Average total EGFP expression levels in 6 cell lines transfected with the mutant pEGFP-C1 vector. The BER deficient colorectal cancer cell line VACO425 has a 2.5 and 2500 fold higher level of total fluorescence than the BER proficient SW480 and HCT116 colorectal cancer cell lines respectively. Error bars represent standard errors for three separate experiments of 5000 cells each (total 15000 cells per cell line). Levels of total EGFP expression are arbitrary units.

6.3.4 Measurement of relative excision repair activity in VACO425 cell line

Relative fluorescence was calculated as outlined in 6.2.5, as (corrected $N_M I_M$)/(corrected $N_{WT} I_{WT}$). 'Relative fluorescence' expresses fluorescence from the mutant vector as a percentage of fluorescence from the wild-type vector. Figure 6.7 shows that in the VACO425 cell line, the mutant vector was expressed at 0.1% of the level seen in the wild-type vector.

From the uncorrected raw data, the average fluorescence intensity of a VACO425 cell transfected with the wild type vector ($I_{Average}$) was 391 units. In the 15,000 VACO425 cells transfected with the wild-type vector, 559 cells had a level of fluorescence intensity above the 10^2 unit cut-off used to determine total EGFP expression from the wild-type vector (see 6.2.5), this is termed N_{WT}. The uncorrected total level of fluorescence from the WT vector was therefore N_{WT} x $I_{Average}$, or 218,569 units. The uncorrected total level of fluorescence from the mutant vector is therefore 0.1% of the WT level, or 219 units. This is approximately equivalent to 1 cell with a level of fluorescence $>10^2$ units, or 0.56 cells with a level of fluorescence equivalent to $I_{Average}$ (391 units). Therefore, assuming similar transfection efficiencies from the wild-type and mutant vectors, we have observed between 1 in 559 and 1 in 1000 vectors harbouring a mutation that leads to EGFP expression.

Data presented previously in Chapter 3 for the VACO425 cell line reveal that across the *APC*, *TGFβR2* and *GPD1* genes, a total of 127 mutations were detected in a total of 151.148 Kb sequence analysed, corresponding to a general mutation frequency of 0.84 mutations/Kb. The pEGFP-C1 vector is 4.731 Kb, if we assume that it too is subject to a mutation frequency of 0.84 mutations/Kb, we would expect ~ 4 mutations per vector when analysed using our cloning and sequencing approach. If we assume that these 4 mutations occur at random along the sequence of the vector, then 1 in 1184 vectors would be mutated at the critical G residue of codon 45 (4731 / 4). If the G were mutated to a T or a C, then expression of EGFP would be restored, if G were mutated to an A, codon 45 would still code for a stop codon. Therefore,

based on a mutation frequency of 0.84 mutations/Kb we would expect 1 in 1775 vectors to harbour a mutation that leads to EGFP expression.

The observed frequency of EGFP expression from the mutant vector is between 1.7 and 3 fold greater than expected to arise by chance due to the mutation frequency observed previously in the VACO425 cells. This suggests that our assay is indeed able to detect the increase in G:C-T:A transversions expected in the presence of BER deficiency, in colorectal cancer cell lines.

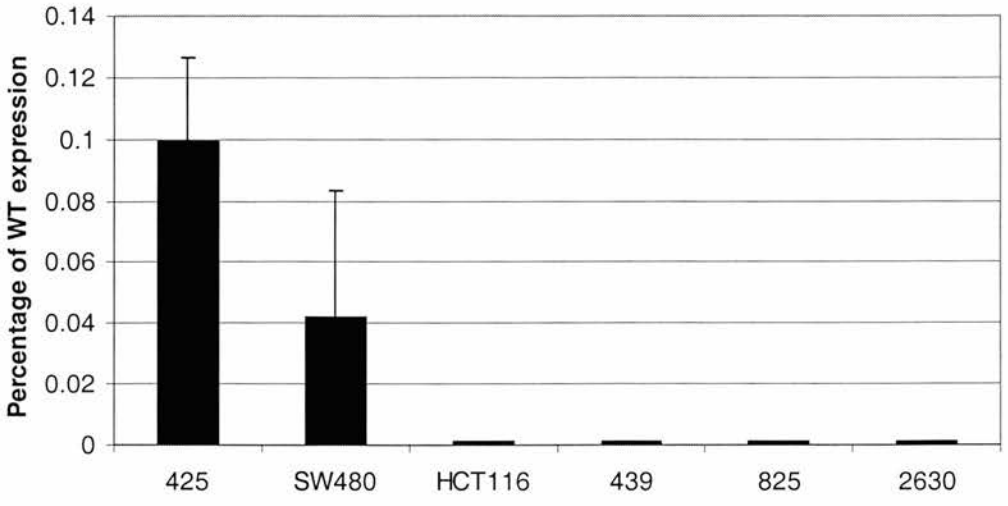

Figure 6.7 Expression of the mutant pEGFP-C1 vector in each cell line is represented as a percentage of the expression seen from the wild-type pEGFP-C1 vector. Expression of EGFP from the mutant pEGFP-C1 vector in the BER deficient colorectal cancer cell line VACO425 is 0.1% of the level detected from the wild-type pEGFP-C1 vector. In the BER proficient SW480 cell line mutant expression is 0.04% of that seen from the wild-type vector. In the other 4 cell lines, including the BER null 2630 cell line, no EGFP expression is detectable from the mutant vector. Error bars show the standard errors of the mean. Levels of total EGFP expression are arbitrary units.

6.4 Discussion

Data presented in this chapter demonstrates that by using a mutant EGFP reporter gene, our assay is able to detect a BER deficiency present in live colorectal cancer cells.

Precise analysis of BER activity in the genomic context would be ideal, however, this is hampered by a lack of techniques that can reliably introduce a single lesion at a defined position on chromosomal DNA. Therefore, the use of plasmids has been developed to probe the pathway indirectly. Our assay aims to measure oxidative DNA damage that remains un-repaired due to deficient BER. The ability of the assay to detect differences in repair activity between different cell types is therefore dependent upon endogenous oxidative damage. In contrast to the assay described in this chapter, a number of previously developed assays for BER activity depend upon the repair of damage that is artificially introduced into the plasmid. Importantly, this has provided evidence that damaged plasmid DNA is processed in the same way as endogenous DNA (Le Page *et al.*, 1998;Le Page *et al.*, 2000), however, these assays have used resistance to restriction enzymes as a non-quantitative method of determining if BER repair of introduced 8-OxoG lesions has occurred. The assay described by Sattler *et al.* introduced an 8-OxoG lesion coupled to a stop codon within the EGFP gene, but is specific for LP-BER, so is unsuitable for assaying total BER activity, as was the aim here.

The introduction of a TGA stop codon containing an 8-OxoG residue in place of a normal G, into the EGFP gene might have allowed us to bypass the initial step in our assay that is dependent upon endogenous oxidative damage taking place. This would produce a more sensitive assay, at the expense of simulating the real situation. Once replication had taken place, mis-pairing of an A with the inserted 8-OxoG lesion (assuming that repair had not occurred) would result in EGFP expression. Another alternative would be to introduce an 8-OxoG lesion directly onto the transcribed (template) strand, rather than onto the untranscribed (coding) strand, that would not disrupt EGFP expression i.e. opposite a TCA serine residue or a TAC tyrosine

residue. Following replication, and in the absence of BER, a G:C-T:A transversion would result in a stop codon (TAA in both cases), and EGFP expression would be halted. However, reports indicate that during transcription RNA Polymerase II stalls at 8-OxoG lesions, but is also able to bypass them, and may insert either a correct C or an incorrect A residue in its place to ensure transcription continues (Kuraoka *et al.*, 2003). For the purposes of our assay, the introduction of 8-OxoG onto the transcribed (template) strand rather than the untranscribed (coding) strand would therefore not be useful, as any transcription taking place prior to replication of the vector, may result in the introduction of a stop codon, even in the presence of functional BER, due to mis-reading of 8-OxoG by RNA Polymerase II. Additionally, there are likely to be more technical difficulties associated with detecting rare events that result in loss of EGFP expression rather than rare events resulting in gain of EGFP expression.

During the development of this assay, one of our aims was that it should be able to analyse BER activity in a variety of different cell types, this is, in theory possible, however, we have shown that the assay is very prone to differences in cell growth and transfection efficiency. This is due to its dependency upon vector replication following endogenous oxidative DNA damage. During the optimisation stage, up to 50% of cells could be seen to have some level of fluorescence from the wild-type vector, 24 hours after transfection (6.3.1). Many less cells had levels $>10^2$ units of fluorescence when analysed by flow cytometry 72 hours after transfection, indicating that a number of cells had lost the vector following replication. Although we have detected a significant increase in the fluorescence levels of the mutant vector in the VACO425 CRC cell line compared with the other cell lines, our stringent cut-off for fluorescence levels ($>10^2$ units) coupled with the loss of the vector from a large number of cells 72 hours after transfection is likely to be under estimating the rate at which G:C-T:A transversions are occurring in BER deficient CRC cells. Using non-replicating plasmids to assay for repair kinetics, Le Page *et al.* have shown that the repair of 8-OxoG:C lesions takes place within 12 hours of transfection (Le Page *et al.*, 1998) suggesting that in our rapidly dividing CRC cell lines it may be possible to reduce the incubation time between transfection and flow cytometry analysis. This

may help to reduce the number of cells that have lost the vector containing the EGFP gene.

The effect of a BER defect causing EGFP expression from the mutant vector appears to be specific to CRC cells, as it is not detected in the BER null 2630 fibroblast cell line. This may be due to the higher proliferation rate in CRC cells compared with fibroblast cells. Firstly, high cell proliferation activity is correlated with cellular endocytotic activity, and transfection will therefore be more efficient in highly proliferating cells (Nakase *et al.*, 2005), whilst this was corrected for using β-galactosidase enzyme activity, the more cells that contain the mutant vector, the more likely it is that an effect from deficient BER will be detected. Additionally, the chance of a BER defect being observed in highly proliferating cells may also be more likely. Reactive oxygen species are generated endogenously by cellular metabolism, therefore, cells that have a higher rate of cellular metabolism, such as tumour cells, may generate more endogenous oxidative damage (Brozmanova *et al.*, 2001). For further optimisation of the assay it will be useful to investigate alternative transfection strategies that may not be so sensitive to cell proliferation rates. These include physical methods such as electroporation and sonoporation. A high rate of cell proliferation may go some way to explaining the results from the SW480 cell line, where we observed a detectable, but non-significant level of fluorescence from the mutant vector in the absence of a BER defect. However, despite the high level of wild-type EGFP expression observed in the HCT116 cell line, which was not effectively corrected for by β-galactosidase enzyme activity, no fluorescence was observed from the mutant vector, even when it is, presumably, present in a high proportion of cells, providing good evidence that expression of the EGFP gene, from the mutant pEGFP-C1 vector is not detected in the absence of a BER defect.

The colon is believed to be inherently more prone to cancer, due to rapid cell division, and exposure to DNA (including oxidative) damage via dietary toxins (Leedham *et al.*, 2005). This goes some way to explaining why patients with MYH associated polyposis, where MYH is bi-allelically inactivated, develop colorectal cancer specifically, rather than cancers in other tissues, i.e. fibroblast cells. As well

as the origin of the cells, the nature of the BER defects in the cells studied in this chapter could be influencing how they perform in our assay. The 2630 fibroblast cell line has been shown to harbour bi-allelic inactivating mutations of *MUTYH*, and whilst the direct effects of these mutations on this cell line have not yet been determined specifically, a study that included a lymphoblastoid cell line with the same G382D/Y165C genotype revealed significantly reduced expression of MYH and reduced levels of 8-OxoG:A repair (Parker *et al.*, 2005a). Likewise, the VACO425 CRC cell line has been characterised as having a 20 fold increase in genomic 8-OxoG, and a 15 fold reduction in repair activity of 8-OxoG:A, compared with SW480 cells. This is presumed to be due to inactivation of *MUTYH*, as reduced levels of mRNA and protein were detected compared with control cell lines, although specific *MUTYH* mutations have not been determined (Parker *et al.*, 2002).

The tumour suppressor gene *p53*, is involved in several DNA repair pathways, either via transactivation of p53-dependent genes or by direct involvement with the p53 protein (Offer *et al.*, 1999;Offer *et al.*, 2001). Whilst the HCT116 CRC cell line is wild-type for *p53,* SW480 cells have been shown to harbour a G-A mutation at codon 273 resulting in an Arginine to Histidine substitution (Liu and Bodmer, 2006). The effect of a variety of *p53* mutations, including the SW480 273R/H mutation were determined by Offer *et al.* in p53 knock-out cells. Along with deletions of the C-terminus, the 273R/H mutation failed to augment BER activity when compared with the wild-type p53 protein. BER activity was therefore concluded to be dependent on a wild-type core domain (amino acids 100 to 300), and an intact C-terminus domain (amino acids 363 to 393), but not an intact N-terminal domain. These findings have implications for our assay, as the detection of fluorescence from the mutant vector in SW480 cells, whilst not statistically significant, may signify a BER deficiency in these cells caused by abrogation of p53 function. Further work is warranted to determine if these cells do indeed contain a BER defect, despite being characterised as largely wild-type for genomic 8-OxoG levels, 8-OxoG:A repair, and levels of MYH mRNA and protein (Parker *et al.*, 2002). Initial western blot analysis has been carried out, and reveals WT levels of p53 protein in the SW480 cell line (data not shown). This is not unexpected, as the 273R/H mutation affects p53's

ability to augment BER, rather than decreasing overall levels of the protein. Interestingly however, significantly reduced levels of p53 were observed in both the VACO425 and HCT116 cell lines (data not shown). The p53 protein has been reported at WT levels in HCT116 (Liu and Bodmer, 2006) so this finding suggests that our cell line may have acquired additional pathogenic mutations or inactivating events. However, the results from our EGFP assay do not suggest that reduced levels of p53 are affecting detectable BER activity. Reduced levels of p53 protein in the VACO425 cell line may provide another means, alongside absent MYH, by which these cells are deficient for BER.

In this chapter we have begun to develop an assay that is able to detect BER deficiency in colorectal cancer cell lines. At this preliminary stage we have obtained some exciting results that indicate that the occurrence of G:C-T:A transversion mutations in cells with a known BER defect may be as much as three times greater than would be expected from the underlying mutation frequency of the cells. To further improve the assay it will be important to test a wider variety of cell lines known to have BER defects, to see if it is possible to replicate what is seen in VACO425, and determine if the assay will be useful for other cell types. Another important aspect of the assay that requires further investigation is the manipulation of the physiological conditions of the cell lines. We were unsuccessful with initial H_2O_2 treatments since they were toxic to the cells, however sublethal concentrations of other reactive oxygen species (such as superoxide, 6-hydroxydopamine and heme) should be investigated. Increasing oxidative damage within cells should increase the number of mutant vectors that express EGFP when BER is deficient, improving the reliability of the assay. A reasonably straightforward proof-of-principle experiment would be to co-transfect the VACO425 cell line with a vector containing the mutant *EGFP* gene alongside a WT *MUTYH* gene. Transfection of *MUTYH* deficient cell lines with exogenous WT *MUTYH* has partially corrected MYH activity in a previous study (Parker *et al.*, 2005b). Therefore, if the vector containing WT *MUTYH* abolishes fluorescence from the mutant *EGFP* gene, there would be strong evidence that fluorescence is directly due to dysfunctional MYH protein, resulting from bi-allelic *MUTYH* mutations. Lastly, it would be interesting to trial vectors with

an 8-OxoG lesion directly inserted, this should reduce the complexity of the assay, as it would no longer rely on endogenous DNA damage, so should be more efficient at detecting BER activity.

Chapter 7

Discussion and Summary

The main aim of this thesis was to investigate our hypothesis of inherent instability in the *APC* gene. This was approached in a number of different ways, including attempting to determine the underlying frequency with which mutations arise in the *APC* gene in both the *in vivo* (patient samples) and *in vitro* (cell lines) situations. Subsequently, the mutations that are ultimately selected for during the tumourigenic process were also determined. Alongside this comparison of mutation frequencies, types and spectra, *in silico* techniques were used to establish how inherent instability of *APC* might contribute to tumour initiation and progression.

As cancers with DNA repair defects influence genetic alterations, elucidation of how they exert this effect may provide new insight into the contribution of DNA repair defects in colon carcinogenesis. Therefore, throughout this project the influence that DNA repair plays on the *APC* gene, and how the presence or absence of DNA repair affects mutations in individual cells and genetic alterations in whole tumours has been investigated.

Lastly, with the aim of ascertaining the effect of loosing DNA repair on cellular mutation frequency, a novel assay has begun to be optimised.

7.1 Mutation frequency at the *APC* gene in normal lymphoblastoid cell lines

Work in this thesis has utilised two normal lymphoblastoid cell lines, from patients devoid of cancer and without a family history of cancer. As they are derived from non-tumourous tissue, these cells are not subject to selection pressures for mutations that provide the cell with a growth advantage, and which result in successive waves of clonal evolution. Whilst these cell lines have been EBV transformed, previous work from the lab has indicated that this is unlikely to have an effect on mutation frequency (Bacon *et al.*, 2001f). In these normal cells, factors biasing the observable mutation frequency (such as mutations in DNA repair genes) are minimised, and therefore it was proposed that the presence of inherent instability at given sequences would be effectively "unmasked". It was believed that because these cell lines are not cancer derived, any detection of excess mutations would indicate that the *APC* gene sequence is inherently unstable, adding further strength to support the importance of an inherent mutation rate in tumourigenesis (Loeb, 2001a).

In both normal control cell lines (ConC and ConA) there were no significant differences between the mutation frequencies observed in the two control genes (*TGFβR2* and *GPD1*), suggesting that this low level of mutation (0-0.2 mutations/Kb) represents a basal level with which mutations arise in most genes. Previous estimations of the underlying mutation frequency of the genome have been estimated using the *HPRT* reporter gene. In T-lymphocytes the mutation frequency of this 43Kb gene has been estimated to be ~5 x 10^{-6} in healthy 20 year old subjects (Robinson *et al.*, 1994). This equates to ~2.15 x 10^{-4} mutations/Kb, approximately 1000 times less than the estimation from our assay. Even though the method used to determine mutation frequency at the *HPRT* reporter gene requires the selection of individual cells that have lost HPRT function, which is likely to result in an underestimation of the true frequency with which mutations arise, this comparison suggests that our assay is grossly over-estimating mutation frequency.

As discussed in section 3.4, the expansion of mutant clones within a population of cells would result in an estimation of mutation frequency that is highly variable and include many overestimates (Lea and Coulson, 1949). However, the results detected in this thesis, using a PCR/cloning assay are remarkably consistent from clone to clone. This observation, alongside the high frequency of mutations observed in all cell lines is well explained by the bulk of mutations having arisen during PCR amplification. Therefore, it is not possible to conclude that instability of the *APC* gene has been uncovered in these normal lymphoblastoid cell lines.

Two areas of future work, important for more detailed investigation of the possible instability in the *APC* gene include the examination of the normal *in vivo* situation. At this time ethical approval has not been obtained for the collection of colonic mucosa from healthy individuals without a family history of colorectal cancer. The ability to determine whether normal colonic mucosa behaves in a similar way to normal lymphoblastoid cells will be advantageous, although it should be considered that instability at the *APC* gene may well be masked by tissue specific selective pressures in such samples.

Hypermutability of the *APC* gene sequence would provide a mechanism to explain the high frequency of mutations that occur early in the carcinogenic process. However, the elevated frequencies of mutations that have been observed are likely to be due to the PCR/cloning assay used, and therefore the data do not support this hypothesis. Infact, the data are consistent with the *APC* gene being normal with respect to mutagenesis and the observed mutation frequency being due to selection pressure acting on cells with advantageous mutations, due to the key gatekeeper role of *APC*.

7.2 *In silico* analysis of *APC* mutation frequency

The *in silico* work presented in Chapter 5 was carried out with the aim of providing a greater insight into the proposed instability of the *APC* gene. Previous studies have used the footprint of mutations left behind by a defective DNA polymerase or repair system to identify the source of mutation (Kuraguchi *et al.*, 2000c;Al Tassan *et al.*, 2002a;Shao *et al.*, 2004b). However, similar work carried out here has not been able to offer a viable and constant alternative explanation to explain the frequency of *APC* mutations seen in CRC. At this stage three possible explanations include inherent instability of the *APC* gene sequence that may act alongside clonal selection, undetected mutational mechanisms and selection pressure acting on *APC* allowing advantageous mutations to become fixed. Evidence to support the theory of instability in *APC* is highlighted by the high value of K_S obtained from the human-mouse alignment *in silico* work, suggesting that the *APC* gene is likely to be subject to a high neutral mutation rate.

There are two main scenarios from which inherent instability of the *APC* gene sequence could arise, both of which warrant further investigation. Firstly, the *APC* gene may be located in a faster mutating region of the genome compared with the two control genes analysed, or secondly, it might be that it is the gene sequence itself that is susceptible to mutation.

Despite extensive analysis of *de novo* mutations, we have so far been unable to identify any aspect of the *APC* gene sequence that might make it more susceptible to mutation, although further investigation into the role played by polymerase δ in contributing to *de novo* mutations will be important (Ripa *et al.*, 2002e).

If the *APC* gene is in fact located in a fast mutating region of the genome then it might be the conformation, or local environment of the DNA that is affecting mutation frequency. Additional features, such as the structure and/or conformation of the DNA surrounding *APC* may lead to the differential mutation frequencies observed, as they can provide different exposure to mutagenic agents or DNA

processing enzymes, and their features warrant further investigation. A similar hypothesis has previously been proposed, suggesting that chromatin structure may affect the susceptibility of regions of DNA to mutation (Hsu, 1975). Although, whether genes in open or closed chromatin are the most susceptible to mutation is still debated (Gazave *et al.*, 2005). The chromatin structure of the DNA surrounding the three genes studied here has been determined (James Prendergast, personal communication). The *APC* gene overlaps a clone in very closed chromatin, *TGFβR2* is in marginally closed chromatin, and *GPD1* is believed to be in average chromatin, based on the nearest available clones (Gilbert *et al.*, 2005). Therefore, the chromatin environment surrounding *APC* is not significantly different to the two control genes. At this stage, with the role played by chromatin environment in susceptibility to mutation still under debate, it is not possible to determine accurately if this is contributing to the increase in *APC* mutation frequency that we have observed. Nevertheless, it does underline the importance of considering factors such as DNA conformation and environment.

7.3 DNA repair mechanisms may influence the spectrum of mutations arising in the *APC* gene

The data presented in this thesis on the analysis of *APC* 'instability' in a variety of DNA repair backgrounds using our 'snap-shot' technique produced a large amount of information. As has been discussed previously in Chapter 3, much of this data appears likely to have arisen as a result of PCR error, however some of the mutation spectrum data may well be pointing towards a novel influence of defective DNA repair in the *APC* gene and across the genome.

The use of cell lines lbl-1260 and lbl-1261, derived from normal tissue, but with deficient MMR, has previously revealed mutations that specifically arise as a result of absent MMR, but with a reduced bias of selection pressures and clonal advantage that tumourigenic cells are subject to (Bacon *et al.*, 2001g). For this reason we have used these cell lines to investigate the relationship between MMR and *APC* instability, and directly compare the results with those obtained in the two control cell lines. However, due to the non-repetitive sequence that was analysed in this study it is likely that alongside mutation frequencies being potentially influenced by the expansion of mutant clones in cell culture, errors from PCR have also had a large impact on our results.

It was expected that selection pressures present in colorectal cancer cell lines would bias the observable mutation frequency, making the dissection of events dependent solely on *APC* instability more difficult in the HCT116 MMR defective colorectal cancer cell line. The trend in mutation frequencies and spectra across all three genes studied in this cell line were indistinguishable from that in the MMR defective lymphoblastoid cell line, lbl-1261. Also, our two BER defective cell lines, although being of different origin (fibroblast and colorectal cancer) gave indistinguishable results across all three genes. This evidence strongly suggests that the observed mutation frequency and spectra were indeed greatly influenced by factors other than instability in the *APC* gene sequence.

In tumour samples a greater proportion of transversions is observed than in normal mucosa samples. We have considered the possibility that these transversion mutations may be appearing above the 'noise' of transition mutations (possibly caused by PCR error), and may represent true mutations occurring in the tumour DNA. We have not been able to identify a possible mechanism to explain how inactive *MUTYH* causes an increase in general transversion mutations. However, an increase in transversion mutations arising, compared with transition mutations, significantly increases the likelihood of a coding amino acid becoming a STOP codon. In heterozygote *MUTYH* tumours the increase in transversion mutations compared with sporadic tumours suggest a distinct mechanism by which these individuals are predisposed to colorectal cancer, and may provide insights to the mechanism of tumourigenesis in these cancers, and also supports the evidence of a heterozygous effect in susceptibility to the disease. Currently, the contribution, or otherwise, made by heterozygous *MUTYH* mutations to CRC susceptibility is still under debate (Croitoru *et al.*, 2004a;Farrington *et al.*, 2005d;Tenesa *et al.*, 2006c;Jenkins *et al.*, 2006c). Our work suggests that whilst the characteristic inactivating G:C-T:A transversion mutations observed in the *APC* and *KRAS* genes of MAP individuals are not selected for in the same way in heterozygotes (Chapter 4), the presence of a single *MUTYH* mutation appears to influence their tumour DNA, causing it to contain more transversion mutations than that of tumours without any *MUTYH* mutations. This alteration of the mutation spectrum may point towards a mechanism by which the susceptibility of *MUTYH* heterozygotes to developing CRC is increased over individuals with completely wild-type *MUTYH*. It will be interesting to look in more detail at the interactions of MYH with other proteins, to deduce if any of these could be responsible for the *MUTYH* heterozygous phenomenon observed.

Maley *et al.* suggest that an increase in clonal diversity predicts incipient transformation (Maley *et al.*, 2006e). Nevertheless, using our assay normal mucosa samples did not generally have lower levels of diversity than tumour samples, as would be expected. If a truly normal *in vivo* sample had been available to us, it would have been interesting to see if reduced cellular diversity in truly normal

samples could in fact be detected using our assay. It seems that this would be unlikely however, because if the vast differences in mutation frequency observed between *in vivo* samples (7.3.2) were due to clonal expansion in some tumours reducing the diversity in the tumour tissue compared to the normal mucosal sample, this would have produced evidence to support the efficacy of our assay, and this was rarely seen.

The majority of mutations identified in all samples were transitions. A propensity for transition mutations in individual cells from cell lines and *in vivo* samples with MMR gene mutations, has been reported previously in T cells derived from mice (Shao *et al.*, 2002;Shao *et al.*, 2004c). At the *Aprt* reporter gene, all single base substitutions from both an $Mlh1^{-/-}$ and a $Pms2^{-/-}$ mouse were transition mutations, inline with our findings. However, the distribution of transition mutations observed was considerably different to what has been observed in this thesis. Again, this indicates that our assay may be influenced by PCR error. In addition, in the studies of Shao *et al.* no increased mutational load was detected in $Mlh1^{+/-}$ T cells at the *Aprt* reporter gene, despite an increase being detected in $Mlh1^{-/-}$ cells. Our assay clearly indicates that MMR heterozygote cells can have an increased mutational load at three different gene loci.

7.4 Further characterisation of the influence played by DNA repair mechanisms is important for the understanding of tumourigenesis

What has become clear during the work undertaken for this thesis is that an understanding of DNA repair is of fundamental importance for unravelling the complex process of carcinogenesis. To this end, I have begun to develop a more reliable molecular tool for clinical and research purposes that accurately measures DNA BER.

At the present time, despite a variety of tools being available to measure BER, they tend to be technically difficult, time consuming and/or expensive. A number use whole cell extracts, rather than live cells, that are either incubated with an antibody of interest (Parker *et al.*, 2002g;Alhopuro *et al.*, 2005a), or used to measure uracil DNA-glycosylase activity (Collini *et al.*, 2005). Live cell assays tend to require complicated vectors (Carreau *et al.*, 1995b). Whilst an assay does already exist that utilises EGFP expression (Sattler *et al.*, 2003b), only LP-BER, and not SP-BER could be measured. Therefore, there is clearly a need for simple and reliable approaches to compare and monitor BER capacities in various types of cells, under various physiological and pathological conditions.

As discussed in detail in section 6.4, the development of an assay that could be used in a clinical setting was undertaken. During the optimisation process a detectable increase in mutation rate was observed in BER null cells. This assay, although in its early stages of development has the potential to be used in high-throughput screens of cells from a variety of origins.

7.5 Summary

Work presented in this thesis has focused on determining the frequency with which mutations spontaneously arise in the *APC* gene. Central to this has been investigating the role played by DNA repair mechanisms in influencing *APC* mutation frequency and spectrum.

In two wild-type lymphoblastoid cell lines an elevation in mutation frequency in a region of *APC* exon 15 was detected (Chapter 3), whilst this may have indicated inherent instability of gene sequences even when DNA repair mechanisms are intact, further work is required to address the problems in our experimental approach.

Polymorphism and evolutionary data (Chapter 5), shows that the *APC* gene is subject to a high neutral mutation rate and purifying selection. However, the effect of selection acting to remove non-synonymous changes is likely to be modest. In addition, analysis of the type and location of *de novo APC* mutations has provided no alternative mechanism to explain how these mutations arise.

The influence of DNA repair mechanisms on *APC* mutation frequency and spectrum has been investigated using a significant number of cell lines and patient samples. There is an effect on the spectrum of mutations observed in the tumour samples of patients carrying *MUTYH* mutations (Chapter 3). Furthermore, the influence of bi-allelic inactivation of *MUTYH* can be clearly seen on *APC* and *KRAS* somatic inactivating mutations in the Scottish population (Chapter 4).

To further characterise the influence of DNA repair mechanisms in carcinogenesis a novel technique has begun to be developed for directly measuring and comparing the effect of BER deficiency on the mutation frequency of live cells (Chapter 6).

In conclusion, data presented here are largely consistent with the *APC* gene being normal with respect to mutagenesis. However, the contribution played by levels of DNA repair deficiency on the type of mutations that arise adds further complexity to

the situation of tumour formation. This research has added to the understanding of how cancers develop, and the role that DNA repair plays.

Chapter 8

Bibliography

Aarnio,M., Sankila,R., Pukkala,E., Salovaara,R., Aaltonen,L.A., de la,C.A., Peltomaki,P., Mecklin,J.P., Jarvinen,H.J. (1999). Cancer risk in mutation carriers of DNA-mismatch-repair genes. Int.J.Cancer *81*, 214-218.

Acharya,S., Wilson,T., Gradia,S., Kane,M.F., Guerrette,S., Marsischky,G.T., Kolodner,R., Fishel,R. (1996). hMSH2 forms specific mispair-binding complexes with hMSH3 and hMSH6. Proc.Natl.Acad.Sci.U.S.A *93*, 13629-13634.

Al Tassan,N., Chmiel,N.H., Maynard,J., Fleming,N., Livingston,A.L., Williams,G.T., Hodges,A.K., Davies,D.R., David,S.S., Sampson,J.R., Cheadle,J.P. (2002). Inherited variants of MYH associated with somatic G:C-->T:A mutations in colorectal tumors. Nat.Genet. *30*, 227-232.

Al-Tassan,N., Eisen,T., Maynard,J., Bridle,H., Shah,B., Fleischmann,C., Sampson,J.R., Cheadle,J.P., Houlston,R.S. (2004). Inherited variants in MYH are unlikely to contribute to the risk of lung carcinoma. Hum.Genet. *114*, 207-210.

Albuquerque,C., Breukel,C., van der,L.R., Fidalgo,P., Lage,P., Slors,F.J., Leitao,C.N., Fodde,R., Smits,R. (2002). The 'just-right' signaling model: APC somatic mutations are selected based on a specific level of activation of the beta-catenin signaling cascade. Hum.Mol.Genet. *11*, 1549-1560.

Alhopuro,P., Parker,A.R., Lehtonen,R., Enholm,S., Jarvinen,H.J., Mecklin,J.P., Karhu,A., Eshleman,J.R., Aaltonen,L.A. (2005). A novel functionally deficient MYH variant in individuals with colorectal adenomatous polyposis. Hum.Mutat. *26*, 393.

Ameri,A., Fairbanks,V.F., Yanik,G.A., Mahdi,F., Thibodeau,S.N., McCormick,D.J., Boxer,L.A., McDonagh,K.T. (1999). Identification of the molecular genetic defect of patients with methemoglobin M-Kankakee (M-Iwate), alpha87 (F8) His --> Tyr: evidence for an electrostatic model of alphaM hemoglobin assembly. Blood *94*, 1825-1826.

Arber,N., Hibshoosh,H., Moss,S.F., Sutter,T., Zhang,Y., Begg,M., Wang,S., Weinstein,I.B., Holt,P.R. (1996). Increased expression of cyclin D1 is an early event in multistage colorectal carcinogenesis. Gastroenterology *110*, 669-674.

Aretz,S., Uhlhaas,S., Caspari,R., Mangold,E., Pagenstecher,C., Propping,P., Friedl,W. (2004). Frequency and parental origin of de novo APC mutations in familial adenomatous polyposis. Eur.J.Hum.Genet. *12*, 52-58.

Aretz,S., Uhlhaas,S., Goergens,H., Siberg,K., Vogel,M., Pagenstecher,C., Mangold,E., Caspari,R., Propping,P., Friedl,W. (2006a). MUTYH-associated polyposis: 70 of 71 patients with biallelic mutations present with an attenuated or atypical phenotype. Int.J.Cancer *119*, 807-814.

Aretz,S., Uhlhaas,S., Goergens,H., Siberg,K., Vogel,M., Pagenstecher,C., Mangold,E., Caspari,R., Propping,P., Friedl,W. (2006b). MUTYH-associated polyposis: 70 of 71 patients with biallelic mutations present with an attenuated or atypical phenotype. Int.J.Cancer *119*, 807-814.

Bacon, A. L. Investigation of the Mutational Consequences of Defective Mismatch Repair in Humans. 1-7-2001. University of Edinburgh.
Ref Type: Thesis/Dissertation

Bacon,A.L., Farrington,S.M., Dunlop,M.G. (2000). Sequence interruptions confer differential stability at microsatellite alleles in mismatch repair-deficient cells. Hum.Mol.Genet. *9*, 2707-2713.

Bacon,A.L., Farrington,S.M., Dunlop,M.G. (2001). Mutation frequency in coding and non-coding repeat sequences in mismatch repair deficient cells derived from normal human tissue. Oncogene *20*, 7464-7471.

Baeg,G.H., Matsumine,A., Kuroda,T., Bhattacharjee,R.N., Miyashiro,I., Toyoshima,K., Akiyama,T. (1995). The tumour suppressor gene product APC blocks cell cycle progression from G0/G1 to S phase. EMBO J. *14*, 5618-5625.

Barnetson,R.A., Tenesa,A., Farrington,S.M., Nicholl,I.D., Cetnarskyj,R., Porteous,M.E., Campbell,H., Dunlop,M.G. (2006). Identification and survival of carriers of mutations in DNA mismatch-repair genes in colon cancer. N.Engl.J.Med. *354*, 2751-2763.

Batlle,E., Bacani,J., Begthel,H., Jonkheer,S., Gregorieff,A., van de,B.M., Malats,N., Sancho,E., Boon,E., Pawson,T., Gallinger,S., Pals,S., Clevers,H. (2005). EphB receptor activity suppresses colorectal cancer progression. Nature *435*, 1126-1130.

Beroud,C., Soussi,T. (1996). APC gene: database of germline and somatic mutations in human tumors and cell lines. Nucleic Acids Res. *24*, 121-124.

Bertario,L., Russo,A., Sala,P., Varesco,L., Giarola,M., Mondini,P., Pierotti,M., Spinelli,P., Radice,P. (2003). Multiple approach to the exploration of genotype-phenotype correlations in familial adenomatous polyposis. J.Clin.Oncol. *21*, 1698-1707.

Bisgaard,M.L., Fenger,K., Bulow,S., Niebuhr,E., Mohr,J. (1994). Familial adenomatous polyposis (FAP): frequency, penetrance, and mutation rate. Hum.Mutat. *3*, 121-125.

Blaker,H., Scholten,M., Sutter,C., Otto,H.F., Penzel,R. (2003). Somatic mutations in familial adenomatous polyps. Nuclear translocation of beta-catenin requires more than biallelic APC inactivation. Am.J.Clin.Pathol. *120*, 418-423.

Bonelli,L., Martines,H., Conio,M., Bruzzi,P., Aste,H. (1988). Family history of colorectal cancer as a risk factor for benign and malignant tumours of the large bowel. A case-control study. Int.J.Cancer *41*, 513-517.

Bright-Thomas,R.M., Hargest,R. (2003). APC, beta-Catenin and hTCF-4; an unholy trinity in the genesis of colorectal cancer. Eur.J.Surg.Oncol. *29*, 107-117.

Brozmanova,J., Dudas,A., Henriques,J.A. (2001). Repair of oxidative DNA damage--an important factor reducing cancer risk. Minireview. Neoplasma *48*, 85-93.

Bulow,S. (1991). Diagnosis of familial adenomatous polyposis. World J.Surg. *15*, 41-46.

Bustamante,C.D., Fledel-Alon,A., Williamson,S., Nielsen,R., Hubisz,M.T., Glanowski,S., Tanenbaum,D.M., White,T.J., Sninsky,J.J., Hernandez,R.D., Civello,D., Adams,M.D., Cargill,M., Clark,A.G. (2005). Natural selection on protein-coding genes in the human genome. Nature *437*, 1153-1157.

Cahill,D.P., Kinzler,K.W., Vogelstein,B., Lengauer,C. (1999). Genetic instability and darwinian selection in tumours. Trends Cell Biol. *9*, M57-M60.

Cahill,D.P., Lengauer,C., Yu,J., Riggins,G.J., Willson,J.K., Markowitz,S.D., Kinzler,K.W., Vogelstein,B. (1998). Mutations of mitotic checkpoint genes in human cancers. Nature *392*, 300-303.

Calabrese,P., Tavare,S., Shibata,D. (2004a). Pretumor progression: clonal evolution of human stem cell populations. Am.J.Pathol. *164*, 1337-1346.

Calabrese,P., Tsao,J.L., Yatabe,Y., Salovaara,R., Mecklin,J.P., Jarvinen,H.J., Aaltonen,L.A., Tavare,S., Shibata,D. (2004b). Colorectal pretumor progression before and after loss of DNA mismatch repair. Am.J.Pathol. *164*, 1447-1453.

Capella,G., Cronauer-Mitra,S., Pienado,M.A., Perucho,M. (1991). Frequency and spectrum of mutations at codons 12 and 13 of the c-K-ras gene in human tumors. Environ.Health Perspect. *93*, 125-131.

Cardoso,J., Molenaar,L., de Menezes,R.X., van Leerdam,M., Rosenberg,C., Moslein,G., Sampson,J., Morreau,H., Boer,J.M., Fodde,R. (2006). Chromosomal Instability in MYH- and APC-Mutant Adenomatous Polyps. Cancer Res. *66*, 2514-2519.

Carreau,M., Eveno,E., Quilliet,X., Chevalier-Lagente,O., Benoit,A., Tanganelli,B., Stefanini,M., Vermeulen,W., Hoeijmakers,J.H., Sarasin,A., . (1995). Development of a new easy complementation assay for DNA repair deficient human syndromes using cloned repair genes. Carcinogenesis *16*, 1003-1009.

Caspari,R., Friedl,W., Mandl,M., Moslein,G., Kadmon,M., Knapp,M., Jacobasch,K.H., Ecker,K.W., Kreissler-Haag,D., Timmermanns,G., . (1994). Familial adenomatous polyposis: mutation at codon 1309 and early onset of colon cancer. Lancet *343*, 629-632.

Charames,G.S., Bapat,B. (2003). Genomic instability and cancer. Curr.Mol.Med. *3*, 589-596.

Cheadle,J.P., Krawczak,M., Thomas,M.W., Hodges,A.K., Al Tassan,N., Fleming,N., Sampson,J.R. (2002). Different combinations of biallelic APC mutation confer different growth advantages in colorectal tumours. Cancer Res. *62*, 363-366.

Cheadle,J.P., Sampson,J.R. (2003). Exposing the MYtH about base excision repair and human inherited disease. Hum.Mol.Genet. *12 Spec No 2*, R159-R165.

Chuang,J.H., Li,H. (2004). Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome. PLoS.Biol. *2*, E29.

Clark,A.G., Glanowski,S., Nielsen,R., Thomas,P.D., Kejariwal,A., Todd,M.A., Tanenbaum,D.M., Civello,D., Lu,F., Murphy,B., Ferriera,S., Wang,G., Zheng,X., White,T.J., Sninsky,J.J., Adams,M.D., Cargill,M. (2003). Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science *302*, 1960-1963.

Clevers,H., Batlle,E. (2006). EphB/EphrinB receptors and Wnt signaling in colorectal cancer. Cancer Res. *66*, 2-5.

Cohen,M.M., Jr., Kreiborg,S., Lammer,E.J., Cordero,J.F., Mastroiacovo,P., Erickson,J.D., Roeper,P., Martinez-Frias,M.L. (1992). Birth prevalence study of the Apert syndrome. Am.J.Med.Genet. *42*, 655-659.

Coleman,W.B., Tsongalis,G.J. (1999). The role of genomic instability in human carcinogenesis. Anticancer Res. *19*, 4645-4664.

Collini,M., Caccia,M., Chirico,G., Barone,F., Dogliotti,E., Mazzei,F. (2005). Two-photon fluorescence cross-correlation spectroscopy as a potential tool for high-throughput screening of DNA repair activity. Nucleic Acids Res. *33*, e165.

Colussi,C., Parlanti,E., Degan,P., Aquilina,G., Barnes,D., Macpherson,P., Karran,P., Crescenzi,M., Dogliotti,E., Bignami,M. (2002). The mammalian mismatch repair pathway removes DNA 8-oxodGMP incorporated from the oxidized dNTP pool. Curr.Biol. *12*, 912-918.

Croitoru,M.E., Cleary,S.P., Di,N.N., Manno,M., Selander,T., Aronson,M., Redston,M., Cotterchio,M., Knight,J., Gryfe,R., Gallinger,S. (2004a). Association between biallelic and monoallelic germline MYH gene mutations and colorectal cancer risk. J.Natl.Cancer Inst. *96*, 1631-1634.

Croitoru,M.E., Cleary,S.P., Di,N.N., Manno,M., Selander,T., Aronson,M., Redston,M., Cotterchio,M., Knight,J., Gryfe,R., Gallinger,S. (2004b). Association between biallelic and monoallelic germline MYH gene mutations and colorectal cancer risk. J.Natl.Cancer Inst. *96*, 1631-1634.

Crowe FW, Schull WJ, Neel JV (1956). A Clinical. Pathological and Genetic Study of Multiple Neurofibromatosis. Springfield IL: Charles C Thomas.

Davies,H., Bignell,G.R., Cox,C., Stephens,P., Edkins,S., Clegg,S., Teague,J., Woffendin,H., Garnett,M.J., Bottomley,W., Davis,N., Dicks,E., Ewing,R., Floyd,Y., Gray,K., Hall,S., Hawes,R., Hughes,J., Kosmidou,V., Menzies,A., Mould,C., Parker,A., Stevens,C., Watt,S., Hooper,S., Wilson,R., Jayatilake,H., Gusterson,B.A., Cooper,C., Shipley,J., Hargrave,D., Pritchard-Jones,K., Maitland,N., Chenevix-Trench,G., Riggins,G.J., Bigner,D.D., Palmieri,G., Cossu,A., Flanagan,A., Nicholson,A., Ho,J.W., Leung,S.Y., Yuen,S.T., Weber,B.L., Seigler,H.F., Darrow,T.L., Paterson,H., Marais,R., Marshall,C.J., Wooster,R., Stratton,M.R., Futreal,P.A. (2002). Mutations of the BRAF gene in human cancer. Nature *417*, 949-954.

De Vos,M., Hayward,B.E., Picton,S., Sheridan,E., Bonthron,D.T. (2004). Novel PMS2 pseudogenes can conceal recessive mutations causing a distinctive childhood cancer syndrome. Am.J.Hum.Genet. *74*, 954-964.

de Wind,N., Dekker,M., Claij,N., Jansen,L., van Klink,Y., Radman,M., Riggins,G., van,d., V, van't Wout,K., te,R.H. (1999). HNPCC-like cancer predisposition in mice through simultaneous loss of Msh3 and Msh6 mismatch-repair protein functions. Nat.Genet. *23*, 359-362.

di Pietro,M., Bellver,J.S., Menigatti,M., Bannwart,F., Schnider,A., Russell,A., Truninger,K., Jiricny,J., Marra,G. (2005). Defective DNA mismatch repair determines a characteristic transcriptional profile in proximal colon cancers. Gastroenterology *129*, 1047-1059.

Diergaarde,B., van Geloof,W.L., van Muijen,G.N., Kok,F.J., Kampman,E. (2003). Dietary factors and the occurrence of truncating APC mutations in sporadic colon carcinomas: a Dutch population-based study. Carcinogenesis *24*, 283-290.

Dietrich,W.F., Lander,E.S., Smith,J.S., Moser,A.R., Gould,K.A., Luongo,C., Borenstein,N., Dove,W. (1993). Genetic identification of Mom-1, a major modifier locus affecting Min-induced intestinal neoplasia in the mouse. Cell *75*, 631-639.

Dikovskaya,D., Newton,I.P., Nathke,I.S. (2004). The adenomatous polyposis coli protein is required for the formation of robust spindles formed in CSF Xenopus extracts. Mol.Biol.Cell *15*, 2978-2991.

Dunlop,M.G. (1997). Colorectal cancer. BMJ *314*, 1882-1885.

Dunlop,M.G., Farrington,S.M., Bubb,V.J., Cunningham,C., Wright,M., Curtis,L.J., Butt,Z.A., Wright,E., Fleck,B.W., Redhead,D., Mitchell,R., Rainey,J.B., Macintyre,I.M., Carter,D.C., Wyllie,A.H. (1996). Extracolonic features of familial adenomatous polyposis in patients with sporadic colorectal cancer. Br.J.Cancer *74*, 1789-1795.

Eberhart,C.E., Coffey,R.J., Radhika,A., Giardiello,F.M., Ferrenbach,S., DuBois,R.N. (1994). Up-regulation of cyclooxygenase 2 gene expression in human colorectal adenomas and adenocarcinomas. Gastroenterology *107*, 1183-1188.

Eccles,D., Harvey,J., Bateman,A., Ross,F. (2001). A novel 3' mutation in the APC gene in a family presenting with a desmoid tumour. J.Med.Genet. *38*, 861-863.

Ellegren,H., Smith,N.G., Webster,M.T. (2003). Mutation rate variation in the mammalian genome. Curr.Opin.Genet.Dev. *13*, 562-568.

Esteller,M., Sparks,A., Toyota,M., Sanchez-Cespedes,M., Capella,G., Peinado,M.A., Gonzalez,S., Tarafa,G., Sidransky,D., Meltzer,S.J., Baylin,S.B., Herman,J.G. (2000). Analysis of adenomatous polyposis coli promoter hypermethylation in human cancer. Cancer Res. *60*, 4366-4371.

Farrington,S.M., Dunlop,M.G. (1999). Mosaicism and sporadic familial adenomatous polyposis. Am.J.Hum.Genet. *64*, 653-658.

Farrington,S.M., Dunlop,M.G. (2004). Familial Colon Cancer Syndromes and Their Genetics. In: Genetic Predisposition to Cancer.

Farrington,S.M., McKinley,A.J., Carothers,A.D., Cunningham,C., Bubb,V.J., Sharp,L., Wyllie,A.H., Dunlop,M.G. (2002). Evidence for an age-related influence of microsatellite instability on colorectal cancer survival. Int.J.Cancer *98*, 844-850.

Farrington,S.M., Tenesa,A., Barnetson,R., Wiltshire,A., Prendergast,J., Porteous,M., Campbell,H., Dunlop,M.G. (2005). Germline susceptibility to colorectal cancer due to base-excision repair gene defects. Am.J.Hum.Genet. *77*, 112-119.

Fearnhead,N.S., Wilding,J.L., Bodmer,W.F. (2002). Genetics of colorectal cancer: hereditary aspects and overview of colorectal tumorigenesis. Br.Med.Bull. *64*, 27-43.

Fishel,R. (2001). The selection for mismatch repair defects in hereditary nonpolyposis colorectal cancer: revising the mutator hypothesis. Cancer Res. *61*, 7369-7374.

Flores-Rozas,H., Kolodner,R.D. (1998). The Saccharomyces cerevisiae MLH3 gene functions in MSH3-dependent suppression of frameshift mutations. Proc.Natl.Acad.Sci.U.S.A *95*, 12404-12409.

Fodde,R. (2003). The multiple functions of tumour suppressors: it's all in APC. Nat.Cell Biol. *5*, 190-192.

Fodde,R., Kuipers,J., Rosenberg,C., Smits,R., Kielman,M., Gaspar,C., van Es,J.H., Breukel,C., Wiegant,J., Giles,R.H., Clevers,H. (2001a). Mutations in the APC tumour suppressor gene cause chromosomal instability. Nat.Cell Biol. *3*, 433-438.

Fodde,R., Smits,R. (2001). Disease model: familial adenomatous polyposis. Trends Mol.Med. *7*, 369-373.

Fodde,R., Smits,R., Clevers,H. (2001b). APC, signal transduction and genetic instability in colorectal cancer. Nat.Rev.Cancer *1*, 55-67.

Fortini,P., Pascucci,B., Parlanti,E., D'Errico,M., Simonelli,V., Dogliotti,E. (2003). 8-Oxoguanine DNA damage: at the crossroad of alternative repair pathways. Mutat.Res. *531*, 127-139.

Frei,J.V. (1990). Stable genes. Med.Hypotheses *32*, 243-244.

Frei,J.V. (1995). Stable and unstable genes: distribution among species and range of mutations. Med.Hypotheses *44*, 101-102.

Friedl,W., Meuschel,S., Caspari,R., Lamberti,C., Krieger,S., Sengteller,M., Propping,P. (1996). Attenuated familial adenomatous polyposis due to a mutation in the 3' part of the APC gene. A clue for understanding the function of the APC protein. Hum.Genet. *97*, 579-584.

Galio,L., Bouquet,C., Brooks,P. (1999). ATP hydrolysis-dependent formation of a dynamic ternary nucleoprotein complex with MutS and MutL. Nucleic Acids Res. *27*, 2325-2331.

Gao,L., Zhang,J. (2003). Why are some human disease-associated mutations fixed in mice? Trends Genet. *19*, 678-681.

Gatof,D., Ahnen,D. (2002). Primary prevention of colorectal cancer: diet and drugs. Gastroenterol.Clin.North Am. *31*, 587-623, xi.

Gazave,E., Gautier,P., Gilchrist,S., Bickmore,W.A. (2005). Does radial nuclear organisation influence DNA damage? Chromosome.Res. *13*, 377-388.

Geisler,J.P., Goodheart,M.J., Sood,A.K., Holmes,R.J., Hatterman-Zogg,M.A., Buller,R.E. (2003). Mismatch repair gene expression defects contribute to microsatellite instability in ovarian carcinoma. Cancer *98*, 2199-2206.

Gertig,D.M., Hunter,D.J. (1998). Genes and environment in the etiology of colorectal cancer. Semin.Cancer Biol. *8*, 285-298.

Giardiello,F.M., Hamilton,S.R., Krush,A.J., Piantadosi,S., Hylind,L.M., Celano,P., Booker,S.V., Robinson,C.R., Offerhaus,G.J. (1993). Treatment of colonic and rectal adenomas with sulindac in familial adenomatous polyposis. N.Engl.J.Med. *328*, 1313-1316.

Giaretti,W., Venesio,T., Prevosto,C., Lombardo,F., Ceccarelli,J., Molinu,S., Risio,M. (2004). Chromosomal instability and APC gene mutations in human sporadic colorectal adenomas. J.Pathol. *204*, 193-199.

Gilbert,N., Gilchrist,S., Bickmore,W.A. (2005). Chromatin organization in the mammalian nucleus. Int.Rev.Cytol. *242*, 283-336.

Grady,W.M. (2004). Genomic instability and colon cancer. Cancer Metastasis Rev. *23*, 11-27.

Groden,J., Thliveris,A., Samowitz,W., Carlson,M., Gelbert,L., Albertsen,H., Joslyn,G., Stevens,J., Spirio,L., Robertson,M., . (1991). Identification and characterization of the familial adenomatous polyposis coli gene. Cell *66*, 589-600.

Gu,Y., Parker,A., Wilson,T.M., Bai,H., Chang,D.Y., Lu,A.L. (2002). Human MutY homolog, a DNA glycosylase involved in base excision repair, physically and functionally interacts with mismatch repair proteins human MutS homolog 2/human MutS homolog 6. J.Biol.Chem. *277*, 11135-11142.

Halford,S.E., Rowan,A.J., Lipton,L., Sieber,O.M., Pack,K., Thomas,H.J., Hodgson,S.V., Bodmer,W.F., Tomlinson,I.P. (2003). Germline mutations but not somatic changes at the MYH locus contribute to the pathogenesis of unselected colorectal cancers. Am.J.Pathol. *162*, 1545-1548.

Hamilton,S.R., Liu,B., Parsons,R.E., Papadopoulos,N., Jen,J., Powell,S.M., Krush,A.J., Berk,T., Cohen,Z., Tetu,B., . (1995). The molecular basis of Turcot's syndrome. N.Engl.J.Med. *332*, 839-847.

Harfe,B.D., Jinks-Robertson,S. (2000). DNA mismatch repair and genetic instability. Annu.Rev.Genet. *34*, 359-399.

Hayashi,H., Tominaga,Y., Hirano,S., McKenna,A.E., Nakabeppu,Y., Matsumoto,Y. (2002). Replication-associated repair of adenine:8-oxoguanine mispairs by MYH. Curr.Biol. *12*, 335-339.

Hegan,D.C., Narayanan,L., Jirik,F.R., Edelmann,W., Liskay,R.M., Glazer,P.M. (2006). Differing patterns of genetic instability in mice deficient in the mismatch repair genes Pms2, Mlh1, Msh2, Msh3, and Msh6. Carcinogenesis.

Heinen,C.D., Schmutte,C., Fishel,R. (2002). DNA repair and tumorigenesis: lessons from hereditary cancer syndromes. Cancer Biol.Ther. *1*, 477-485.

Heppner,G.K., Trzepacz,C., Tuohy,T.M., Groden,J. (2002). Attenuated APC alleles produce functional protein from internal translation initiation. Proc.Natl.Acad.Sci.U.S.A *99*, 8161-8166.

Hienonen,T., Sammalkorpi,H., Enholm,S., Alhopuro,P., Barber,T.D., Lehtonen,R., Nupponen,N.N., Lehtonen,H., Salovaara,R., Mecklin,J.P., Jarvinen,H., Koistinen,R., Arango,D., Launonen,V., Vogelstein,B., Karhu,A., Aaltonen,L.A. (2005). Mutations in two short noncoding mononucleotide repeats in most microsatellite-unstable colorectal cancers. Cancer Res. *65*, 4607-4613.

Horii,A., Nakatsuru,S., Ichii,S., Nagase,H., Nakamura,Y. (1993). Multiple forms of the APC gene transcripts and their tissue-specific expression. Hum.Mol.Genet. *2*, 283-287.

Hsieh,J.S., Lin,S.R., Chang,M.Y., Chen,F.M., Lu,C.Y., Huang,T.J., Huang,Y.S., Huang,C.J., Wang,J.Y. (2005). APC, K-ras, and p53 gene mutations in colorectal cancer patients: correlation to clinicopathologic features and postoperative surveillance. Am.Surg. *71*, 336-343.

Hsu,T.C. (1975). A possible function of constitutive heterochromatin: the bodyguard hypothesis. Genetics *79 Suppl*, 137-150.

Huang,J., Papadopoulos,N., McKinley,A.J., Farrington,S.M., Curtis,L.J., Wyllie,A.H., Zheng,S., Willson,J.K., Markowitz,S.D., Morin,P., Kinzler,K.W., Vogelstein,B., Dunlop,M.G. (1996). APC mutations in colorectal tumors with mismatch repair deficiency. Proc.Natl.Acad.Sci.U.S.A *93*, 9049-9054.

Huang,J., Zheng,S., Jin,S.H., Zhang,S.Z. (2004). Somatic mutations of APC gene in carcinomas from hereditary non-polyposis colorectal cancer patients. World J.Gastroenterol. *10*, 834-836.

Hurst,L.D. (2002). The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet. *18*, 486.

Hurst,L.D., Smith,N.G. (1999). Do essential genes evolve slowly? Curr.Biol. *9*, 747-750.

Ilyas,M., Straub,J., Tomlinson,I.P., Bodmer,W.F. (1999). Genetic pathways in colorectal and other cancers. Eur.J.Cancer *35*, 335-351.

Issa,J.P. (2004b). CpG island methylator phenotype in cancer. Nat.Rev.Cancer *4*, 988-993.

Issa,J.P. (2004a). CpG island methylator phenotype in cancer. Nat.Rev.Cancer *4*, 988-993.

Jass,J.R. (2000). Familial colorectal cancer: pathology and molecular characteristics. Lancet Oncol. *1*, 220-226.

Jenkins,M.A., Croitoru,M.E., Monga,N., Cleary,S.P., Cotterchio,M., Hopper,J.L., Gallinger,S. (2006). Risk of colorectal cancer in monoallelic and biallelic carriers of MYH mutations: a population-based case-family study. Cancer Epidemiol.Biomarkers Prev. *15*, 312-314.

Jones,S., Emmerson,P., Maynard,J., Best,J.M., Jordan,S., Williams,G.T., Sampson,J.R., Cheadle,J.P. (2002). Biallelic germline mutations in MYH predispose to multiple colorectal adenoma and somatic G:C-->T:A mutations. Hum.Mol.Genet. *11*, 2961-2967.

Jones,S., Lambert,S., Williams,G.T., Best,J.M., Sampson,J.R., Cheadle,J.P. (2004). Increased frequency of the k-ras G12C mutation in MYH polyposis colorectal adenomas. Br.J.Cancer *90*, 1591-1593.

Kambara,T., Whitehall,V.L., Spring,K.J., Barker,M.A., Arnold,S., Wynter,C.V., Matsubara,N., Tanaka,N., Young,J.P., Leggett,B.A., Jass,J.R. (2004). Role of inherited defects of MYH in the development of sporadic colorectal cancer. Genes Chromosomes.Cancer *40*, 1-9.

Kaplan,K.B., Burds,A.A., Swedlow,J.R., Bekir,S.S., Sorger,P.K., Nathke,I.S. (2001). A role for the Adenomatous Polyposis Coli protein in chromosome segregation. Nat.Cell Biol. *3*, 429-432.

Kim,N.G., Rhee,H., Li,L.S., Kim,H., Lee,J.S., Kim,J.H., Kim,N.K., Kim,H. (2002). Identification of MARCKS, FLJ11383 and TAF1B as putative novel target genes in colorectal carcinomas with microsatellite instability. Oncogene *21*, 5081-5087.

Kinzler,K.W., Vogelstein,B. (1996). Lessons from hereditary colorectal cancer. Cell *87*, 159-170.

Klungland,A., Lindahl,T. (1997). Second pathway for completion of human DNA base excision-repair: reconstitution with purified proteins and requirement for DNase IV (FEN1). EMBO J. *16*, 3341-3348.

Kohl,N.E., Kanda,N., Schreck,R.R., Bruns,G., Latt,S.A., Gilbert,F., Alt,F.W. (1983). Transposition and amplification of oncogene-related sequences in human neuroblastomas. Cell *35*, 359-367.

Kolodner,R. (1996). Biochemistry and genetics of eukaryotic mismatch repair. Genes Dev. *10*, 1433-1442.

Kolodner,R.D., Marsischky,G.T. (1999). Eukaryotic DNA mismatch repair. Curr.Opin.Genet.Dev. *9*, 89-96.

Komarova,N.L., Wang,L. (2004). Initiation of Colorectal Cancer: Where do the Two Hits Hit? Cell Cycle *3*.

Kondo,Y., Issa,J.P. (2004). Epigenetic changes in colorectal cancer. Cancer Metastasis Rev. *23*, 29-39.

Kuraguchi,M., Edelmann,W., Yang,K., Lipkin,M., Kucherlapati,R., Brown,A.M. (2000). Tumor-associated Apc mutations in Mlh1-/- Apc1638N mice reveal a mutational signature of Mlh1 deficiency. Oncogene *19*, 5755-5763.

Kuraoka,I., Endou,M., Yamaguchi,Y., Wada,T., Handa,H., Tanaka,K. (2003). Effects of endogenous DNA base lesions on transcription elongation by mammalian RNA polymerase II. Implications for transcription-coupled DNA repair and transcriptional mutagenesis. J.Biol.Chem. *278*, 7294-7299.

Laken,S.J., Petersen,G.M., Gruber,S.B., Oddoux,C., Ostrer,H., Giardiello,F.M., Hamilton,S.R., Hampel,H., Markowitz,A., Klimstra,D., Jhanwar,S., Winawer,S., Offit,K., Luce,M.C., Kinzler,K.W., Vogelstein,B. (1997). Familial colorectal cancer in Ashkenazim due to a hypermutable tract in APC. Nat.Genet. *17*, 79-83.

Lamlum,H., Al Tassan,N., Jaeger,E., Frayling,I., Sieber,O., Reza,F.B., Eckert,M., Rowan,A., Barclay,E., Atkin,W., Williams,C., Gilbert,J., Cheadle,J., Bell,J., Houlston,R., Bodmer,W., Sampson,J., Tomlinson,I. (2000a). Germline APC variants in patients with multiple colorectal adenomas, with evidence for the particular importance of E1317Q. Hum.Mol.Genet. *9*, 2215-2221.

Lamlum,H., Papadopoulou,A., Ilyas,M., Rowan,A., Gillet,C., Hanby,A., Talbot,I., Bodmer,W., Tomlinson,I. (2000b). APC mutations are sufficient for the growth of early colorectal adenomas. Proc.Natl.Acad.Sci.U.S.A *97*, 2225-2228.

Land,H., Parada,L.F., Weinberg,R.A. (1983). Cellular oncogenes and multistep carcinogenesis. Science *222*, 771-778.

Larson,E.D., Iams,K., Drummond,J.T. (2003). Strand-specific processing of 8-oxoguanine by the human mismatch repair pathway: inefficient removal of 8-oxoguanine paired with adenine or cytosine. DNA Repair (Amst) *2*, 1199-1210.

Laurent-Puig,P., Beroud,C., Soussi,T. (1998b). APC gene: database of germline and somatic mutations in human tumors and cell lines. Nucleic Acids Res. *26*, 269-270.

Laurent-Puig,P., Beroud,C., Soussi,T. (1998a). APC gene: database of germline and somatic mutations in human tumors and cell lines. Nucleic Acids Res. *26*, 269-270.

Le Page,F., Guy,A., Cadet,J., Sarasin,A., Gentil,A. (1998). Repair and mutagenic potency of 8-oxoG:A and 8-oxoG:C base pairs in mammalian cells. Nucleic Acids Res. *26*, 1276-1281.

Le Page,F., Klungland,A., Barnes,D.E., Sarasin,A., Boiteux,S. (2000). Transcription coupled repair of 8-oxoguanine in murine cells: the ogg1 protein is required for repair in nontranscribed sequences but not in transcribed sequences. Proc.Natl.Acad.Sci.U.S.A *97*, 8397-8402.

Lea,D.E., Coulson,C.A. (1949). The distribution of the numbers of mutants in bacterial populations. J.Genetics *49*, 264-285.

Leedham,S.J., Schier,S., Thliveris,A.T., Halberg,R.B., Newton,M.A., Wright,N.A. (2005). From gene mutations to tumours - stem cells in gastrointestinal carcinogenesis. Cell Prolif. *38*, 387-405.

Lei,X., Zhu,Y., Tomkinson,A., Sun,L. (2004). Measurement of DNA mismatch repair activity in live cells. Nucleic Acids Res. *32*, e100.

Leibovitz,A., Stinson,J.C., McCombs,W.B., III, McCoy,C.E., Mazur,K.C., Mabry,N.D. (1976). Classification of human colorectal adenocarcinoma cell lines. Cancer Res. *36*, 4562-4569.

Lengauer,C., Kinzler,K.W., Vogelstein,B. (1998). Genetic instabilities in human cancers. Nature *396*, 643-649.

Li,G.M., Modrich,P. (1995). Restoration of mismatch repair to nuclear extracts of H6 colorectal tumor cells by a heterodimer of human MutL homologs. Proc.Natl.Acad.Sci.U.S.A *92*, 1950-1954.

Lichtenstein,P., Holm,N.V., Verkasalo,P.K., Iliadou,A., Kaprio,J., Koskenvuo,M., Pukkala,E., Skytthe,A., Hemminki,K. (2000). Environmental and heritable factors in

the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. N.Engl.J.Med. *343*, 78-85.

Lin,D.P., Wang,Y., Scherer,S.J., Clark,A.B., Yang,K., Avdievich,E., Jin,B., Werling,U., Parris,T., Kurihara,N., Umar,A., Kucherlapati,R., Lipkin,M., Kunkel,T.A., Edelmann,W. (2004). An Msh2 Point Mutation Uncouples DNA Mismatch Repair and Apoptosis. Cancer Res. *64*, 517-522.

Lipkin,S.M., Wang,V., Jacoby,R., Banerjee-Basu,S., Baxevanis,A.D., Lynch,H.T., Elliott,R.M., Collins,F.S. (2000). MLH3: a DNA mismatch repair gene associated with mammalian microsatellite instability. Nat.Genet. *24*, 27-35.

Lipton,L., Halford,S.E., Johnson,V., Novelli,M.R., Jones,A., Cummings,C., Barclay,E., Sieber,O., Sadat,A., Bisgaard,M.L., Hodgson,S.V., Aaltonen,L.A., Thomas,H.J., Tomlinson,I.P. (2003). Carcinogenesis in MYH-associated polyposis follows a distinct genetic pathway. Cancer Res. *63*, 7595-7599.

Liu,B., Nicolaides,N.C., Markowitz,S., Willson,J.K., Parsons,R.E., Jen,J., Papadopolous,N., Peltomaki,P., de la,C.A., Hamilton,S.R., . (1995). Mismatch repair gene defects in sporadic colorectal cancers with microsatellite instability. Nat.Genet. *9*, 48-55.

Liu,Y., Bodmer,W.F. (2006). Analysis of P53 mutations and their expression in 56 colorectal cancer cell lines. Proc.Natl.Acad.Sci.U.S.A *103*, 976-981.

Loeb,L.A. (2001). A mutator phenotype in cancer. Cancer Res. *61*, 3230-3239.

Lothe,R.A., Andersen,S.N., Hofstad,B., Meling,G.I., Peltomaki,P., Heim,S., Brogger,A., Vatn,M., Rognum,T.O., Borresen,A.L. (1995). Deletion of 1p loci and microsatellite instability in colorectal polyps. Genes Chromosomes.Cancer *14*, 182-188.

Lu,S.L., Kawabata,M., Imamura,T., Akiyama,Y., Nomizu,T., Miyazono,K., Yuasa,Y. (1998). HNPCC associated with germline mutation in the TGF-beta type II receptor gene. Nat.Genet. *19*, 17-18.

Lubomierski,N., Plotz,G., Wormek,M., Engels,K., Kriener,S., Trojan,J., Jungling,B., Zeuzem,S., Raedle,J. (2005). BRAF mutations in colorectal carcinoma suggest two entities of microsatellite-unstable tumors. Cancer *104*, 952-961.

Luchtenborg,M., Weijenberg,M.P., Roemen,G.M., De Bruine,A.P., Van Den Brandt,P.A., Lentjes,M.H., Brink,M., Van Engeland,M., Goldbohm,R.A., De Goeij,A.F. (2004). APC mutations in sporadic colorectal carcinomas from the Netherlands cohort study. Carcinogenesis.

Luongo,C., Moser,A.R., Gledhill,S., Dove,W.F. (1994). Loss of Apc+ in intestinal adenomas from Min mice. Cancer Res. *54*, 5947-5952.

Luria,S.E., Delbruck,M. (1943). Mutations of bacteria from virus sensitivity to virus resistance. Genetics *28*, 491-511.

Lynch,H.T., de la Chapelle,A. (1999). Genetic susceptibility to non-polyposis colorectal cancer. J.Med.Genet. *36*, 801-818.

MacPhee,M., Chepenik,K.P., Liddell,R.A., Nelson,K.K., Siracusa,L.D., Buchberg,A.M. (1995). The secretory phospholipase A2 gene is a candidate for the Mom1 locus, a major modifier of ApcMin-induced intestinal neoplasia. Cell *81*, 957-966.

Macpherson,P., Barone,F., Maga,G., Mazzei,F., Karran,P., Bignami,M. (2005). 8-oxoguanine incorporation into DNA repeats in vitro and mismatch recognition by MutSalpha. Nucleic Acids Res. *33*, 5094-5105.

Maher,E.R., Iselius,L., Yates,J.R., Littler,M., Benjamin,C., Harris,R., Sampson,J., Williams,A., Ferguson-Smith,M.A., Morton,N. (1991). Von Hippel-Lindau disease: a genetic study. J.Med.Genet. *28*, 443-447.

Maley,C.C., Galipeau,P.C., Finley,J.C., Wongsurawat,V.J., Li,X., Sanchez,C.A., Paulson,T.G., Blount,P.L., Risques,R.A., Rabinovitch,P.S., Reid,B.J. (2006). Genetic clonal diversity predicts progression to esophageal adenocarcinoma. Nat.Genet. *38*, 468-473.

Markowitz,S., Wang,J., Myeroff,L., Parsons,R., Sun,L., Lutterbaugh,J., Fan,R.S., Zborowska,E., Kinzler,K.W., Vogelstein,B., . (1995). Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability. Science *268*, 1336-1338.

Matsumine,A., Ogai,A., Senda,T., Okumura,N., Satoh,K., Baeg,G.H., Kawahara,T., Kobayashi,S., Okada,M., Toyoshima,K., Akiyama,T. (1996). Binding of APC to the human homolog of the Drosophila discs large tumor suppressor protein. Science *272*, 1020-1023.

Matsumoto,T., Nakamura,S., Esaki,M., Yao,T., Iida,M. (2006). Effect of the non-steroidal anti-inflammatory drug sulindac on colorectal adenomas of uncolectomized familial adenomatous polyposis. J.Gastroenterol.Hepatol. *21*, 251-257.

Meira,L.B., Cheo,D.L., Reis,A.M., Claij,N., Burns,D.K., te,R.H., Friedberg,E.C. (2002). Mice defective in the mismatch repair gene Msh2 show increased predisposition to UVB radiation-induced skin cancer. DNA Repair (Amst) *1*, 929-934.

Michils,G., Tejpar,S., Thoelen,R., van Cutsem,E., Vermeesch,J.R., Fryns,J.P., Legius,E., Matthijs,G. (2005). Large deletions of the APC gene in 15% of mutation-negative patients with classical polyposis (FAP): a Belgian study. Hum.Mutat. *25*, 125-134.

Miyaki,M., Konishi,M., Kikuchi-Yanoshita,R., Enomoto,M., Igari,T., Tanaka,K., Muraoka,M., Takahashi,H., Amada,Y., Fukayama,M., . (1994). Characteristics of somatic mutation of the adenomatous polyposis coli gene in colorectal tumors. Cancer Res. *54*, 3011-3020.

Miyaki,M., Tanaka,K., Kikuchi-Yanoshita,R., Muraoka,M., Konishi,M. (1995). Familial polyposis: recent advances. Crit Rev.Oncol.Hematol. *19*, 1-31.

Miyoshi,Y., Nagase,H., Ando,H., Horii,A., Ichii,S., Nakatsuru,S., Aoki,T., Miki,Y., Mori,T., Nakamura,Y. (1992). Somatic mutations of the APC gene in colorectal tumors: mutation cluster region in the APC gene. Hum.Mol.Genet. *1*, 229-233.

Modrich,P. (1991a). Mechanisms and biological effects of mismatch repair. Annu.Rev.Genet. *25*, 229-253.

Modrich,P. (1991b). Mechanisms and biological effects of mismatch repair. Annu.Rev.Genet. *25*, 229-253.

Modrich,P., Lahue,R. (1996). Mismatch repair in replication fidelity, genetic recombination, and cancer biology. Annu.Rev.Biochem. *65*, 101-133.

Moloney,D.M., Slaney,S.F., Oldridge,M., Wall,S.A., Sahlin,P., Stenman,G., Wilkie,A.O. (1996). Exclusive paternal origin of new mutations in Apert syndrome. Nat.Genet. *13*, 48-53.

Morrin,M., Kelly,M., Barrett,N., Delaney,P. (1994). Mutations of Ki-ras and p53 genes in colorectal cancer and their prognostic significance. Gut *35*, 1627-1631.

Morrison,C., Vagnarelli,P., Sonoda,E., Takeda,S., Earnshaw,W.C. (2003). Sister chromatid cohesion and genome stability in vertebrate cells. Biochem.Soc.Trans. *31*, 263-265.

Moser,A.R., Pitot,H.C., Dove,W.F. (1990). A dominant mutation that predisposes to multiple intestinal neoplasia in the mouse. Science *247*, 322-324.

Muniappan,B.P., Thilly,W.G. (2002). The DNA polymerase beta replication error spectrum in the adenomatous polyposis coli gene contains human colon tumor mutational hotspots. Cancer Res. *62*, 3271-3275.

Nagase,H., Nakamura,Y. (1993). Mutations of the APC (adenomatous polyposis coli) gene. Hum.Mutat. *2*, 425-434.

Nakase,M., Okumura,K., Kamei,T., Nakamura,S., Inui,M., Tagawa,T. (2005). The influence of oral tumor cell proliferation activity and membrane potential on the transfection efficiency of a cationic liposome. Oncol.Rep. *14*, 1487-1491.

Nathke,I. (2005). Relationship between the role of the adenomatous polyposis coli protein in colon cancer and its contribution to cytoskeletal regulation. Biochem.Soc.Trans. *33*, 694-697.

Nathke,I.S. (1999). The adenomatous polyposis coli protein. Mol.Pathol. *52*, 169-173.

Niessen,R.C., Sijmons,R.H., Ou,J., Olthof,S.G., Osinga,J., Ligtenberg,M.J., Hogervorst,F.B., Weiss,M.M., Tops,C.M., Hes,F.J., de Bock,G.H., Buys,C.H.,

Kleibeuker,J.H., Hofstra,R.M. (2006). MUTYH and the mismatch repair system: partners in crime? Hum.Genet. *119*, 206-211.

Oda,S., Zhao,Y., Maehara,Y. (2005). Microsatellite instability in gastrointestinal tract cancers: a brief update. Surg.Today *35*, 1005-1015.

Offer,H., Milyavsky,M., Erez,N., Matas,D., Zurer,I., Harris,C.C., Rotter,V. (2001). Structural and functional involvement of p53 in BER in vitro and in vivo. Oncogene *20*, 581-589.

Offer,H., Wolkowicz,R., Matas,D., Blumenstein,S., Livneh,Z., Rotter,V. (1999). Direct involvement of p53 in the base excision repair pathway of the DNA repair machinery. FEBS Lett. *450*, 197-204.

Olschwang,S., Hamelin,R., Laurent-Puig,P., Thuille,B., De Rycke,Y., Li,Y.J., Muzeau,F., Girodet,J., Salmon,R.J., Thomas,G. (1997). Alternative genetic pathways in colorectal carcinogenesis. Proc.Natl.Acad.Sci.U.S.A *94*, 12122-12127.

Otsuka,K., Suzuki,T., Shibata,H., Kato,S., Sakayori,M., Shimodaira,H., Kanamaru,R., Ishioka,C. (2003). Analysis of the human APC mutation spectrum in a saccharomyces cerevisiae strain with a mismatch repair defect. Int.J.Cancer *103*, 624-630.

Ouyang,H., Furukawa,T., Abe,T., Kato,Y., Horii,A. (1998). The BAX gene, the promoter of apoptosis, is mutated in genetically unstable cancers of the colorectum, stomach, and endometrium. Clin.Cancer Res. *4*, 1071-1074.

Ouyang,H., Shiwaku,H.O., Hagiwara,H., Miura,K., Abe,T., Kato,Y., Ohtani,H., Shiiba,K., Souza,R.F., Meltzer,S.J., Horii,A. (1997). The insulin-like growth factor II receptor gene is mutated in genetically unstable cancers of the endometrium, stomach, and colorectum. Cancer Res. *57*, 1851-1854.

Palombo,F., Iaccarino,I., Nakajima,E., Ikejima,M., Shimada,T., Jiricny,J. (1996). hMutSbeta, a heterodimer of hMSH2 and hMSH3, binds to insertion/deletion loops in DNA. Curr.Biol. *6*, 1181-1184.

Pangilinan,F., Spencer,F. (1996). Abnormal kinetochore structure activates the spindle assembly checkpoint in budding yeast. Mol.Biol.Cell *7*, 1195-1208.

Papadopoulos,N., Nicolaides,N.C., Wei,Y.F., Ruben,S.M., Carter,K.C., Rosen,C.A., Haseltine,W.A., Fleischmann,R.D., Fraser,C.M., Adams,M.D., . (1994). Mutation of a mutL homolog in hereditary colon cancer. Science *263*, 1625-1629.

Parker,A.R., O'Meally,R.N., Oliver,D.H., Hua,L., Nelson,W.G., DeWeese,T.L., Eshleman,J.R. (2002). 8-Hydroxyguanosine repair is defective in some microsatellite stable colorectal cancer cells. Cancer Res. *62*, 7230-7233.

Parker,A.R., Sieber,O.M., Shi,C., Hua,L., Takao,M., Tomlinson,I.P., Eshleman,J.R. (2005a). Cells with pathogenic biallelic mutations in the human MUTYH gene are defective in DNA damage binding and repair. Carcinogenesis *26*, 2010-2018.

Parker,A.R., Sieber,O.M., Shi,C., Hua,L., Takao,M., Tomlinson,I.P., Eshleman,J.R. (2005b). Cells with pathogenic biallelic mutations in the human MUTYH gene are defective in DNA damage binding and repair. Carcinogenesis *26*, 2010-2018.

Parkin,D.M., Pisani,P., Ferlay,J. (1999). Estimates of the worldwide incidence of 25 major cancers in 1990. Int.J.Cancer *80*, 827-841.

Parsons,R., Li,G.M., Longley,M., Modrich,P., Liu,B., Berk,T., Hamilton,S.R., Kinzler,K.W., Vogelstein,B. (1995). Mismatch repair deficiency in phenotypically normal human cells. Science *268*, 738-740.

Pawlowski,J., Kraft,A.S. (2000). Bax-induced apoptotic cell death. Proc.Natl.Acad.Sci.U.S.A *97*, 529-531.

Peltomaki,P. (2001). Deficient DNA mismatch repair: a common etiologic factor for colon cancer. Hum.Mol.Genet. *10*, 735-740.

Peltomaki,P., Aaltonen,L.A., Sistonen,P., Pylkkanen,L., Mecklin,J.P., Jarvinen,H., Green,J.S., Jass,J.R., Weber,J.L., Leach,F.S., . (1993). Genetic mapping of a locus predisposing to human colorectal cancer. Science *260*, 810-812.

Pierce,A.J., Johnson,R.D., Thompson,L.H., Jasin,M. (1999). XRCC3 promotes homology-directed repair of DNA damage in mammalian cells. Genes Dev. *13*, 2633-2638.

Pisani,P., Parkin,D.M., Bray,F., Ferlay,J. (1999). Erratum: Estimates of the worldwide mortality from 25 cancers in 1990. Int. J. Cancer, 83, 18-29 (1999). Int.J.Cancer *83*, 870-873.

Plug,A.W., Clairmont,C.A., Sapi,E., Ashley,T., Sweasy,J.B. (1997). Evidence for a role for DNA polymerase beta in mammalian meiosis. Proc.Natl.Acad.Sci.U.S.A *94*, 1327-1331.

Powell,S.M., Zilz,N., Beazer-Barclay,Y., Bryan,T.M., Hamilton,S.R., Thibodeau,S.N., Vogelstein,B., Kinzler,K.W. (1992). APC mutations occur early during colorectal tumorigenesis. Nature *359*, 235-237.

Prader,A. (1956). [Adrenogenital syndrome, adrenogenital salt deficiency syndrome and Cushing's syndrome in childhood.]. Schweiz.Med.Wochenschr. *86*, 289-299.

Prolla,T.A., Pang,Q., Alani,E., Kolodner,R.D., Liskay,R.M. (1994). MLH1, PMS1, and MSH2 interactions during the initiation of DNA mismatch repair in yeast. Science *265*, 1091-1093.

Prosser,J., vanHeyningen,V. (1998). PAX6 mutations reviewed. Hum.Mutat. *11*, 93-108.

Rajagopalan,H., Bardelli,A., Lengauer,C., Kinzler,K.W., Vogelstein,B., Velculescu,V.E. (2002). Tumorigenesis: RAF/RAS oncogenes and mismatch-repair status. Nature *418*, 934.

Rampino,N., Yamamoto,H., Ionov,Y., Li,Y., Sawai,H., Reed,J.C., Perucho,M. (1997b). Somatic frameshift mutations in the BAX gene in colon cancers of the microsatellite mutator phenotype. Science *275*, 967-969.

Rampino,N., Yamamoto,H., Ionov,Y., Li,Y., Sawai,H., Reed,J.C., Perucho,M. (1997a). Somatic frameshift mutations in the BAX gene in colon cancers of the microsatellite mutator phenotype. Science *275*, 967-969.

Ripa,R., Bisgaard,M.L., Bulow,S., Nielsen,F.C. (2002). De novo mutations in familial adenomatous polyposis (FAP). Eur.J.Hum.Genet. *10*, 631-637.

Roberts,B.T., Farr,K.A., Hoyt,M.A. (1994). The Saccharomyces cerevisiae checkpoint gene BUB1 encodes a novel protein kinase. Mol.Cell Biol. *14*, 8282-8291.

Robinson,D.R., Goodall,K., Albertini,R.J., O'Neill,J.P., Finette,B., Sala-Trepat,M., Moustacchi,E., Tates,A.D., Beare,D.M., Green,M.H., . (1994). An analysis of in vivo hprt mutant frequency in circulating T-lymphocytes in the normal human population: a comparison of four datasets. Mutat.Res. *313*, 227-247.

Rowan,A.J., Lamlum,H., Ilyas,M., Wheeler,J., Straub,J., Papadopoulou,A., Bicknell,D., Bodmer,W.F., Tomlinson,I.P. (2000). APC mutations in sporadic colorectal tumors: A mutational "hotspot" and interdependence of the "two hits". Proc.Natl.Acad.Sci.U.S.A *97*, 3352-3357.

Roy,H.K., Gulizia,J., DiBaise,J.K., Karolski,W.J., Ansari,S., Madugula,M., Hart,J., Bissonnette,M., Wali,R.K. (2004). Polyethylene glycol inhibits intestinal neoplasia and induces epithelial apoptosis in Apc(min) mice. Cancer Lett. *215*, 35-42.

Russell,A.M., Zhang,J., Luz,J., Hutter,P., Chappuis,P.O., Berthod,C.R., Maillet,P., Mueller,H., Heinimann,K. (2005). Prevalence of MYH germline mutations in Swiss APC mutation-negative polyposis patients. Int.J.Cancer.

Sale,S., Tunstall,R.G., Ruparelia,K.C., Potter,G.A., Steward,W.P., Gescher,A.J. (2005). Comparison of the effects of the chemopreventive agent resveratrol and its synthetic analog trans 3,4,5,4'-tetramethoxystilbene (DMU-212) on adenoma development in the Apc(Min+) mouse and cyclooxygenase-2 in human-derived colon cancer cells. Int.J.Cancer *115*, 194-201.

Samowitz,W.S., Albertsen,H., Herrick,J., Levin,T.R., Sweeney,C., Murtaugh,M.A., Wolff,R.K., Slattery,M.L. (2005). Evaluation of a large, population-based sample supports a CpG island methylator phenotype in colon cancer. Gastroenterology *129*, 837-845.

Sampson,J.R., Dolwani,S., Jones,S., Eccles,D., Ellis,A., Evans,D.G., Frayling,I., Jordan,S., Maher,E.R., Mak,T., Maynard,J., Pigatto,F., Shaw,J., Cheadle,J.P. (2003). Autosomal recessive colorectal adenomatous polyposis due to inherited mutations of MYH. Lancet *362*, 39-41.

Samuelsson,B., Akesson,H.O. (1988). Relative fertility and mutation rate in neurofibromatosis. Hereditas *108*, 169-171.

Sansom,O.J., Zabkiewicz,J., Bishop,S.M., Guy,J., Bird,A., Clarke,A.R. (2003). MBD4 deficiency reduces the apoptotic response to DNA-damaging agents in the murine small intestine. Oncogene *22*, 7130-7136.

Sattler,U., Frit,P., Salles,B., Calsou,P. (2003). Long-patch DNA repair synthesis during base excision repair in mammalian cells. EMBO Rep. *4*, 363-367.

Schofield,M.J., Hsieh,P. (2003). DNA mismatch repair: molecular mechanisms and biological function. Annu.Rev.Microbiol. *57*, 579-608.

Shao,C., Deng,L., Chen,Y., Kucherlapati,R., Stambrook,P.J., Tischfield,J.A. (2004). Mlh1 mediates tissue-specific regulation of mitotic recombination. Oncogene *23*, 9017-9024.

Shao,C., Yin,M., Deng,L., Stambrook,P.J., Doetschman,T., Tischfield,J.A. (2002). Loss of heterozygosity and point mutation at Aprt locus in T cells and fibroblasts of Pms2-/- mice. Oncogene *21*, 2840-2845.

Shibata,D. (2006). Clonal diversity in tumor progression. Nat.Genet. *38*, 402-403.

Shibata,H., Toyama,K., Shioya,H., Ito,M., Hirota,M., Hasegawa,S., Matsumoto,H., Takano,H., Akiyama,T., Toyoshima,K., Kanamaru,R., Kanegae,Y., Saito,I., Nakamura,Y., Shiba,K., Noda,T. (1997). Rapid colorectal adenoma formation initiated by conditional targeting of the Apc gene. Science *278*, 120-123.

Shoemaker,A.R., Luongo,C., Moser,A.R., Marton,L.J., Dove,W.F. (1997). Somatic mutational mechanisms involved in intestinal tumor formation in Min mice. Cancer Res. *57*, 1999-2006.

Sieber,O.M., Howarth,K.M., Thirlwell,C., Rowan,A., Mandir,N., Goodlad,R.A., Gilkar,A., Spencer-Dene,B., Stamp,G., Johnson,V., Silver,A., Yang,H., Miller,J.H., Ilyas,M., Tomlinson,I.P. (2004). Myh deficiency enhances intestinal tumorigenesis in multiple intestinal neoplasia (ApcMin/+) mice. Cancer Res. *64*, 8876-8881.

Sieber,O.M., Lamlum,H., Crabtree,M.D., Rowan,A.J., Barclay,E., Lipton,L., Hodgson,S., Thomas,H.J., Neale,K., Phillips,R.K., Farrington,S.M., Dunlop,M.G., Mueller,H.J., Bisgaard,M.L., Bulow,S., Fidalgo,P., Albuquerque,C., Scarano,M.I., Bodmer,W., Tomlinson,I.P., Heinimann,K. (2002). Whole-gene APC deletions cause classical familial adenomatous polyposis, but not attenuated polyposis or "multiple" colorectal adenomas. Proc.Natl.Acad.Sci.U.S.A *99*, 2954-2958.

Sieber,O.M., Tomlinson,I.P., Lamlum,H. (2000). The adenomatous polyposis coli (APC) tumour suppressor--genetics, function and disease. Mol.Med.Today *6*, 462-469.

Silva,J.C., Kondrashov,A.S. (2002). Patterns in spontaneous mutation revealed by human-baboon sequence comparison. Trends Genet. *18*, 544-547.

Slattery,M.L., Samowitz,W., Ballard,L., Schaffer,D., Leppert,M., Potter,J.D. (2001b). A molecular variant of the APC gene at codon 1822: its association with diet, lifestyle, and risk of colon cancer. Cancer Res. *61*, 1000-1004.

Slattery,M.L., Samowitz,W., Ballard,L., Schaffer,D., Leppert,M., Potter,J.D. (2001a). A molecular variant of the APC gene at codon 1822: its association with diet, lifestyle, and risk of colon cancer. Cancer Res. *61*, 1000-1004.

Slupphaug,G., Kavli,B., Krokan,H.E. (2003). The interacting pathways for prevention and repair of oxidative DNA damage. Mutat.Res. *531*, 231-251.

Smith,N.G., Eyre-Walker,A. (2003). Human disease genes: patterns and predictions. Gene *318*, 169-175.

Smits,R., Kielman,M.F., Breukel,C., Zurcher,C., Neufeld,K., Jagmohan-Changur,S., Hofland,N., van Dijk,J., White,R., Edelmann,W., Kucherlapati,R., Khan,P.M., Fodde,R. (1999). Apc1638T: a mouse model delineating critical domains of the adenomatous polyposis coli protein involved in tumorigenesis and development. Genes Dev. *13*, 1309-1321.

Sohn,K.J., Choi,M., Song,J., Chan,S., Medline,A., Gallinger,S., Kim,Y.I. (2003). Msh2 deficiency enhances somatic Apc and p53 mutations in Apc+/-Msh2-/- mice. Carcinogenesis *24*, 217-224.

Sparks,A.B., Morin,P.J., Vogelstein,B., Kinzler,K.W. (1998). Mutational analysis of the APC/beta-catenin/Tcf pathway in colorectal cancer. Cancer Res. *58*, 1130-1134.

Stamatoyannopoulos,G., Nute,P.E. (1982). De novo mutations producing unstable Hbs or Hbs M. II. Direct estimates of minimum nucleotide mutation rates in man. Hum.Genet. *60*, 181-188.

Stark,L.A., Din,F.V., Zwacka,R.M., Dunlop,M.G. (2001). Aspirin-induced activation of the NF-kappaB signaling pathway: a novel mechanism for aspirin-mediated apoptosis in colon cancer cells. FASEB J. *15*, 1273-1275.

Stark,L.A., Dunlop,M.G. (2005). Nucleolar sequestration of RelA (p65) regulates NF-kappaB-driven transcription and apoptosis. Mol.Cell Biol. *25*, 5985-6004.

Stephens,K., Kayes,L., Riccardi,V.M., Rising,M., Sybert,V.P., Pagon,R.A. (1992). Preferential mutation of the neurofibromatosis type 1 gene in paternally derived chromosomes. Hum.Genet. *88*, 279-282.

Stoler,D.L., Chen,N., Basik,M., Kahlenberg,M.S., Rodriguez-Bigas,M.A., Petrelli,N.J., Anderson,G.R. (1999). The onset and extent of genomic instability in sporadic colorectal tumor progression. Proc.Natl.Acad.Sci.U.S.A *96*, 15121-15126.

Strauss,B.S. (1998). Hypermutability in carcinogenesis. Genetics *148*, 1619-1626.

Su,L.K., Kinzler,K.W., Vogelstein,B., Preisinger,A.C., Moser,A.R., Luongo,C., Gould,K.A., Dove,W.F. (1992). Multiple intestinal neoplasia caused by a mutation in the murine homolog of the APC gene. Science *256*, 668-670.

Su,L.K., Kohlmann,W., Ward,P.A., Lynch,P.M. (2002). Different familial adenomatous polyposis phenotypes resulting from deletions of the entire APC exon 15. Hum.Genet. *111*, 88-95.

Su,L.K., Steinbach,G., Sawyer,J.C., Hindi,M., Ward,P.A., Lynch,P.M. (2000). Genomic rearrangements of the APC tumor-suppressor gene in familial adenomatous polyposis. Hum.Genet. *106*, 101-107.

Su,S.S., Modrich,P. (1986). Escherichia coli mutS-encoded protein binds to mismatched DNA base pairs. Proc.Natl.Acad.Sci.U.S.A *83*, 5057-5061.

Takenoshita,S., Tani,M., Nagashima,M., Hagiwara,K., Bennett,W.P., Yokota,J., Harris,C.C. (1997). Mutation analysis of coding sequences of the entire transforming growth factor beta type II receptor gene in sporadic human colon cancer using genomic DNA and intron primers. Oncogene *14*, 1255-1258.

Tenesa,A., Campbell,H., Barnetson,R., Porteous,M., Dunlop,M., Farrington,S.M. (2006). Association of MUTYH and colorectal cancer. Br.J.Cancer *95*, 239-242.

Thierry Soussi. APC Mutation Database. 2004.
Ref Type: Personal Communication

Tokairin,Y., Kakinuma,S., Arai,M., Nishimura,M., Okamoto,M., Ito,E., Akashi,M., Miki,Y., Kawano,T., Iwai,T., Shimada,Y. (2006). Accelerated growth of intestinal tumours after radiation exposure in Mlh1-knockout mice: evaluation of the late effect of radiation on a mouse model of HNPCC. Int.J.Exp.Pathol. *87*, 89-99.

Tortola,S., Marcuello,E., Gonzalez,I., Reyes,G., Arribas,R., Aiza,G., Sancho,F.J., Peinado,M.A., Capella,G. (1999). p53 and K-ras gene mutations correlate with tumor aggressiveness but are not of routine prognostic value in colorectal cancer. J.Clin.Oncol. *17*, 1375-1381.

Truta,B., Allen,B.A., Conrad,P.G., Weinberg,V., Miller,G.A., Pomponio,R., Lipton,L.R., Guerra,G., Tomlinson,I.P., Sleisenger,M.H., Kim,Y.S., Terdiman,J.P. (2005). A comparison of the phenotype and genotype in adenomatous polyposis patients with and without a family history. Fam.Cancer *4*, 127-133.

Tsuchiya,T., Tamura,G., Sato,K., Endoh,Y., Sakata,K., Jin,Z., Motoyama,T., Usuba,O., Kimura,W., Nishizuka,S., Wilson,K.T., James,S.P., Yin,J., Fleisher,A.S., Zou,T., Silverberg,S.G., Kong,D., Meltzer,S.J. (2000). Distinct methylation patterns of two APC gene promoters in normal and cancerous gastric epithelia. Oncogene *19*, 3642-3646.

Umar,A., Risinger,J.I., Glaab,W.E., Tindall,K.R., Barrett,J.C., Kunkel,T.A. (1998). Functional overlap in mismatch repair by human MSH3 and MSH6. Genetics *148*, 1637-1646.

van der, Houven W. X-ray induced carcinogenesis in Eu-pim-1 and Apc1683N mice. 21-10-0098. Leiden University.
Ref Type: Thesis/Dissertation

Vandrovcova,J., Stekrova,J., Kebrdlova,V., Kohoutova,M. (2004). Molecular analysis of the APC and MYH genes in Czech families affected by FAP or multiple adenomas: 13 novel mutations. Hum.Mutat. *23*, 397.

Veigl,M.L., Kasturi,L., Olechnowicz,J., Ma,A.H., Lutterbaugh,J.D., Periyasamy,S., Li,G.M., Drummond,J., Modrich,P.L., Sedwick,W.D., Markowitz,S.D. (1998). Biallelic inactivation of hMLH1 by epigenetic gene silencing, a novel mechanism causing human MSI cancers. Proc.Natl.Acad.Sci.U.S.A *95*, 8698-8702.

Venesio,T., Balsamo,A., Rondo-Spaudo,M., Varesco,L., Risio,M., Ranzani,G.N. (2003). APC haploinsufficiency, but not CTNNB1 or CDH1 gene mutations, accounts for a fraction of familial adenomatous polyposis patients without APC truncating mutations. Lab Invest *83*, 1859-1866.

Vogelstein,B., Fearon,E.R., Hamilton,S.R., Kern,S.E., Preisinger,A.C., Leppert,M., Nakamura,Y., White,R., Smits,A.M., Bos,J.L. (1988). Genetic alterations during colorectal-tumor development. N.Engl.J.Med. *319*, 525-532.

Vogul,F., Motulsky,A.G. (1986). Human Genetics; Problems and Approaches. Springer-Verlag.

Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M., An,P., Antonarakis,S.E., Attwood,J., Baertsch,R., Bailey,J., Barlow,K., Beck,S., Berry,E., Birren,B., Bloom,T., Bork,P., Botcherby,M., Bray,N., Brent,M.R., Brown,D.G., Brown,S.D., Bult,C., Burton,J., Butler,J., Campbell,R.D., Carninci,P., Cawley,S., Chiaromonte,F., Chinwalla,A.T., Church,D.M., Clamp,M., Clee,C., Collins,F.S., Cook,L.L., Copley,R.R., Coulson,A., Couronne,O., Cuff,J., Curwen,V., Cutts,T., Daly,M., David,R., Davies,J., Delehaunty,K.D., Deri,J., Dermitzakis,E.T., Dewey,C., Dickens,N.J., Diekhans,M., Dodge,S., Dubchak,I., Dunn,D.M., Eddy,S.R., Elnitski,L., Emes,R.D., Eswara,P., Eyras,E., Felsenfeld,A., Fewell,G.A., Flicek,P., Foley,K., Frankel,W.N., Fulton,L.A., Fulton,R.S., Furey,T.S., Gage,D., Gibbs,R.A., Glusman,G., Gnerre,S., Goldman,N., Goodstadt,L., Grafham,D., Graves,T.A., Green,E.D., Gregory,S., Guigo,R., Guyer,M., Hardison,R.C., Haussler,D., Hayashizaki,Y., Hillier,L.W., Hinrichs,A., Hlavina,W., Holzer,T., Hsu,F., Hua,A., Hubbard,T., Hunt,A., Jackson,I., Jaffe,D.B., Johnson,L.S., Jones,M., Jones,T.A., Joy,A., Kamal,M., Karlsson,E.K., Karolchik,D., Kasprzyk,A., Kawai,J., Keibler,E., Kells,C., Kent,W.J., Kirby,A., Kolbe,D.L., Korf,I., Kucherlapati,R.S., Kulbokas,E.J., Kulp,D., Landers,T., Leger,J.P., Leonard,S., Letunic,I., Levine,R., Li,J., Li,M., Lloyd,C., Lucas,S., Ma,B., Maglott,D.R., Mardis,E.R., Matthews,L., Mauceli,E., Mayer,J.H., McCarthy,M., McCombie,W.R., McLaren,S., McLay,K., McPherson,J.D., Meldrim,J., Meredith,B., Mesirov,J.P., Miller,W., Miner,T.L., Mongin,E., Montgomery,K.T., Morgan,M., Mott,R., Mullikin,J.C., Muzny,D.M., Nash,W.E., Nelson,J.O., Nhan,M.N., Nicol,R., Ning,Z., Nusbaum,C., O'Connor,M.J., Okazaki,Y., Oliver,K., Overton-Larty,E., Pachter,L., Parra,G., Pepin,K.H., Peterson,J., Pevzner,P., Plumb,R., Pohl,C.S.,

Poliakov,A., Ponce,T.C., Ponting,C.P., Potter,S., Quail,M., Reymond,A., Roe,B.A., Roskin,K.M., Rubin,E.M., Rust,A.G., Santos,R., Sapojnikov,V., Schultz,B., Schultz,J., Schwartz,M.S., Schwartz,S., Scott,C., Seaman,S., Searle,S., Sharpe,T., Sheridan,A., Shownkeen,R., Sims,S., Singer,J.B., Slater,G., Smit,A., Smith,D.R., Spencer,B., Stabenau,A., Stange-Thomann,N., Sugnet,C., Suyama,M., Tesler,G., Thompson,J., Torrents,D., Trevaskis,E., Tromp,J., Ucla,C., Ureta-Vidal,A., Vinson,J.P., Von Niederhausern,A.C., Wade,C.M., Wall,M., Weber,R.J., Weiss,R.B., Wendl,M.C., West,A.P., Wetterstrand,K., Wheeler,R., Whelan,S., Wierzbowski,J., Willey,D., Williams,S., Wilson,R.K., Winter,E., Worley,K.C., Wyman,D., Yang,S., Yang,S.P., Zdobnov,E.M., Zody,M.C., Lander,E.S. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520-562.

Watson,P., Lynch,H.T. (1993). Extracolonic cancer in hereditary nonpolyposis colorectal cancer. Cancer *71*, 677-685.

Williams,C.S., Mann,M., DuBois,R.N. (1999). The role of cyclooxygenases in inflammation, cancer, and development. Oncogene *18*, 7908-7916.

Woerner,S.M., Kloor,M., Mueller,A., Rueschoff,J., Friedrichs,N., Buettner,R., Buzello,M., Kienle,P., Knaebel,H.P., Kunstmann,E., Pagenstecher,C., Schackert,H.K., Moslein,G., Vogelsang,H., von Knebel,D.M., Gebert,J.F. (2005). Microsatellite instability of selective target genes in HNPCC-associated colon adenomas. Oncogene.

Worm,J., Christensen,C., Gronbaek,K., Tulchinsky,E., Guldberg,P. (2004). Genetic and epigenetic alterations of the APC gene in malignant melanoma. Oncogene *23*, 5215-5226.

Xie,Y., Yang,H., Cunanan,C., Okamoto,K., Shibata,D., Pan,J., Barnes,D.E., Lindahl,T., McIlhatton,M., Fishel,R., Miller,J.H. (2004). Deficiencies in mouse Myh and Ogg1 result in tumor predisposition and G to T mutations in codon 12 of the K-ras oncogene in lung tumors. Cancer Res. *64*, 3096-3102.

Yamada,T., Takaoka,A.S., Naishiro,Y., Hayashi,R., Maruyama,K., Maesawa,C., Ochiai,A., Hirohashi,S. (2000). Transactivation of the multidrug resistance 1 gene by T-cell factor 4/beta-catenin complex in early colorectal carcinogenesis. Cancer Res. *60*, 4761-4766.

Yan,H., Dobbie,Z., Gruber,S.B., Markowitz,S., Romans,K., Giardiello,F.M., Kinzler,K.W., Vogelstein,B. (2002). Small changes in expression affect predisposition to tumorigenesis. Nat.Genet. *30*, 25-26.

Yang,G., Scherer,S.J., Shell,S.S., Yang,K., Kim,M., Lipkin,M., Kucherlapati,R., Kolodner,R.D., Edelmann,W. (2004). Dominant effects of an Msh6 missense mutation on DNA repair and cancer susceptibility. Cancer Cell *6*, 139-150.

Zabner,J., Fasbender,A.J., Moninger,T., Poellinger,K.A., Welsh,M.J. (1995). Cellular and molecular barriers to gene transfer by a cationic lipid. J.Biol.Chem. *270*, 18997-19007.

Zauber,N.P., Sabbath-Solitare,M., Marotta,S.P., Bishop,D.T. (2003). The characterization of somatic APC mutations in colonic adenomas and carcinomas in Ashkenazi Jews with the APC I1307K variant using linkage disequilibrium. J.Pathol. *199*, 146-151.

Zhang,H., Richards,B., Wilson,T., Lloyd,M., Cranston,A., Thorburn,A., Fishel,R., Meuth,M. (1999). Apoptosis induced by overexpression of hMSH2 or hMLH1. Cancer Res. *59*, 3021-3027.

Zhang,J., Lindroos,A., Ollila,S., Russell,A., Marra,G., Mueller,H., Peltomaki,P., Plasilova,M., Heinimann,K. (2006). Gene Conversion Is a Frequent Mechanism of Inactivation of the Wild-Type Allele in Cancers from MLH1/MSH2 Deletion Carriers. Cancer Res. *66*, 659-664.

Zhang,T., Otevrel,T., Gao,Z., Gao,Z., Ehrlich,S.M., Fields,J.Z., Boman,B.M. (2001). Evidence that APC regulates survivin expression: a possible mechanism contributing to the stem cell origin of colon cancer. Cancer Res. *61*, 8664-8667.

Zhou,X., Huang,L. (1994). DNA transfection mediated by cationic liposomes containing lipopolylysine: characterization and mechanism of action. Biochim.Biophys.Acta *1189*, 195-203.

Appendix A

APC Mutations

Cell Line	Position ID	APC Change	nt	Codon	Target Sequence
ConC	E12	A-G	274	827	ApA
	F8	A-G	719	975	ApA
	E6	A-G	729	979	ApA
	E6	A-G	758	988	ApA
	H4	A-G	829	1012	ApA
	C7	A-G	531	913	ApC
	A8	A-G	170	792	ApG
	A5	A-G	355	854	ApG
	B10	A-G	665	957	ApG
	A10	A-G	677	961	ApT
	C2	A-G	839	1015	ApT
	G6	A-G	860	1022	ApT
	C10	C-T	590	932	CpA
	B2	C-T	289	832	CpC
	E8	C-T	804	1004	CpC
	B8	C-T	810	100	CpC
	C6	C-T	363	857	CpG
	A7	C-T	326	844	CpT
	G4	C-T	572	926	CpT
	A9	G-A	432	880	GpA
	B10	G-A	745	984	GpA
	F2	G-A	838	1015	GpA
	F1	G-A	293	833	GpC
	H11	G-A	574	927	GpC
	H2	T-C	181	796	TpA
	F7	T-C	582	930	TpA
	D8	T-C	874	1027	TpA
	B10	T-C	226	811	TpC
	C5	T-C	304	837	TpC
	C10	T-C	493	900	TpC
	C4	T-C	695	967	TpC
	E6	T-C	754	987	TpC
	C3	T-C	756	988	TpG
	C10	T-C	849	1019	TpG
	A3	T-C	316	841	TpT
	B5	T-C	348	852	TpT
	B12	T-C	684	964	TpT
	G2	T-C	685	964	TpT
ConA	65	A-G	58	755	ApG
	53	A-G	58	755	ApG
	14	A-G	62	756	ApA
	46	A-G	78	761	ApG
	28	A-G	125	777	ApC
	4	A-G	174	793	ApA
	56	A-G	264	823	ApC
	5	A-G	307	838	ApG
	6	A-G	335	847	ApA
	27	A-G	401	869	ApT
	13	A-G	505	904	ApG
	5	A-G	545	917	ApT
	26	A-G	721	976	ApG
	32	A-G	735	980	ApC
	27	A-G	913	1040	ApG
	3	C-T	522	909	CpA
	54	G-A	563	923	GpA
	50	G-A	782	996	GpT
	61	T-C	27	744	TpT
	26	T-C	30	745	TpG
	6	T-C	46	751	TpC
	46	T-C	100	769	TpT
	14	T-C	286	831	TpT
	37	T-C	301	836	TpC
	6	T-C	486	897	TpC
	14	T-C	546	917	TpG
	42	T-C	605	937	TpC
	44	T-C	613	940	TpC
	11	T-C	639	948	TpA
	24	T-C	670	959	TpC
	62	T-C	673	960	TpC
	23	T-C	717	974	TpA
	67	T-C	748	985	TpC
	67	T-C	843	1016	TpA
	70	T-C	888	1031	TpT
	58	T-C	888	1031	TpT
	32	T-C	907	1038	TpC
	56	T-C	921	1042	TpC
	69	T-G	93	766	TpC
Ibl-1261	G3	A-G	337	848	ApA
	C6	A-G	339	849	ApG
	E10	A-G	359	855	ApA
	8	A-G	37	748	ApG
	56	A-G	62	756	ApA
	83	A-G	253	820	ApG

TGFBR2 Mutations

Cell Line	Position ID	TGFBR2 Change	nt	Codon	Target Sequence
ConC	E9	A-G	190	202	ApC
	C5	A-G	265	227	ApG
	B9	A-G	265	227	ApG
	C9	A-G	323	246	ApC
	F11	A-G	620	345	ApC
	F11	A-G	620	345	ApC
	F1	A-G	691	369	ApG
	C1	C-T	72	162	CpA
	C4	C-T	567	327	CpG
	E1	C-T	803	406	CpT
	E11	T-C	26	147	TpT
	F3	T-C	605	340	TpC
ConA	A5	A-G	400	272	ApC
	G4	G-A	405	273	GpG
	C5	G-A	744	386	GpA
Ibl-1261	G2	A-G	495	303	ApC
	B1	A-G	530	315	ApG
	F10	A-G	571	329	ApA
	B3	C-T	144	186	CpT
	D7	C-T	294	236	CpA
	G2	C-T	324	246	CpA
	B8	C-T	565	327	CpA
	C4	G-A	832	416	GpG
	B5	T-C	142	186	TpA
	A11	T-C	665	360	TpC
HCT116	A4	C-T	234	216	CpA
	B6	C-T	252	222	CpC
Ibl-1260	G8	A-G	197	204	ApG
	B7	A-G	452	289	ApG
	F9	A-G	610	342	ApG
	F9	A-G	746	387	ApG
	C8	C-T	292	236	CpA
	C8	C-T	312	242	CpA
	F9	C-T	326	247	CpC
	A9	C-T	649	355	CpG
	A10	C-T	720	378	CpC
	C9	G-A	378	264	GpA
	E9	T-C	117	177	TpG
	A7	T-C	856	424	TpA
2630	A4	A-G	69	161	ApG
	H6	A-G	170	195	ApG
	F2	A-G	337	251	ApA
	A2	A-G	353	256	ApG
	G4	A-G	488	301	ApG
	B2	A-G	490	302	ApA
	E5	A-G	587	334	ApG
	E6	A-G	719	378	ApC
	H6	C-T	150	188	CpT
	H6	C-T	411	275	CpA
	B2	G-A	213	209	GpG
	C3	G-A	451	289	GpA
	C3	G-A	469	295	GpA
	G4	G-A	635	350	GpC
	B2	T-C	242	219	TpG
	A1	T-C	268	228	TpC
	E5	T-C	433	283	TpA
	C6	T-C	463	293	TpT
	D2	T-C	843	419	TpT
VACO425	B11	A-G	38	151	ApA
	D3	A-G	53	156	ApT
	G10	A-G	80	165	ApA
	A2	A-G	108	174	ApC
	C8	A-G	302	239	ApG
	C5	A-G	446	287	ApG
	A8	A-G	452	289	ApG
	A1	A-G	515	310	ApG
	G10	A-G	553	323	ApT
	A1	A-G	691	369	ApG
	E1	A-G	749	388	ApC
	B3	C-T	659	358	CpT
	B12	G-A	118	178	GpC
	D9	G-A	254	223	GpC
	B11	G-A	519	311	GpC
	D9	G-A	751	389	GpA
	B3	G-A	761	392	GpC
	E10	T-C	61	159	TpT
	A8	T-C	68	161	TpA
	D8	T-C	242	219	TpG
	D3	T-C	356	257	TpC
	B1	T-C	371	262	TpG
	G1	T-C	665	360	TpC
	G4	T-C	730	382	TpC

Cell Line	Position ID	APC Change	nt	Codon	Target Sequence
Ibl-1261 cont	D10	C-T	49	752	CpT
	31	C-T	265	824	CpC
	49	C-T	421	876	CpG
	12	G-A	934	1047	GpA
	16	G-A	577	928	GpC
	29	G-A	124	777	GpA
	49	G-A	351	852	GpG
	72	G-A	658	955	GpA
	101	G-A	930	1045	GpA
	A10	T-C	25	744	TpC
	G8	T-C	260	822	TpT
	E10	T-C	436	881	TpC
	E8	T-C	539	915	TpG
	F10	T-C	539	915	TpG
	E3	T-C	881	1029	TpT
	F8	T-C	901	1036	TpT
	30	T-C	775	944	TpT
	85	T-C	775	944	TpC
HCT116	84	A-G	140	782	ApG
	84	A-G	161	789	ApG
	59	A-G	360	855	ApC
	28	A-G	419	875	ApG
	86	A-G	447	884	ApG
	3	A-G	69	758	ApG
	10	A-G	250	819	ApT
	22	A-G	719	975	ApA
	33	C-T	437	881	CpC
	94	C-T	483	896	CpA
	61	C-T	820	1009	CpA
	44	C-T	449	885	CpC
	59	C-T	578	928	CpC
	22	C-T	740	982	CpG
	2	G-A	544	917	GpA
	16	T-C	145	784	TpC
	1	T-C	145	784	TpC
	23	T-C	493	900	TpC
	78	T-C	613	940	TpC
	76	T-C	765	990	TpG
Ibl-1260	93	A-G	19	742	ApT
	56	A-G	37	748	ApG
	39	A-G	62	756	ApA
	59	A-G	69	758	ApG
	25	A-G	75	760	ApG
	H3	A-G	78	762	ApG
	59	A-G	78	761	ApG
	C6	A-G	102	770	ApT
	C5	A-G	125	777	ApC
	12	A-G	125	777	ApC
	46	A-G	163	790	ApG
	63	A-G	165	790	ApC
	95	A-G	175	794	ApG
	85	A-G	215	807	ApT
	9	A-G	218	808	ApT
	1	A-G	220	809	ApA
	C1	A-G	241	816	ApC
	93	A-G	253	820	ApC
	E8	A-G	267	825	ApT
	E7	A-G	341	849	ApT
	59	A-G	341	849	ApT
	26	A-G	419	875	ApG
	70	A-G	497	901	ApG
	15	A-G	533	913	ApT
	16	A-G	583	930	ApC
	G3	A-G	631	946	ApC
	17	A-G	700	696	ApG
	50	A-G	703	970	ApG
	65	A-G	707	971	ApT
	57	A-G	803	1003	ApC
	C12	A-G	875	1027	ApT
	10	A-G	913	1040	ApG
	44	C-T	80	762	CpA
	70	C-T	80	762	CpA
	E2	C-T	302	836	CpA
	C3	C-T	315	841	CpT
	H5	C-T	416	874	CpA
	G4	C-T	496	901	CpA
	D12	C-T	557	921	CpA
	F11	C-T	623	943	CpA
	21	C-T	804	1003	CpC
	22	G-A	314	840	GpC
	6	G-A	666	957	GpA
	H5	G-A	692	966	GpC
	C12	G-T	942	1050	GpG
	55	T-C	255	820	TpG

GPD1 Mutations

Cell Line	Position ID	GPD1 Change	nt	Codon	Target Sequence
ConA	G5	A-G	201	236	ApG
	E8	T-C	182	230	TpG
	D11	T-C	264	257	TpG
	E7	T-C	264	257	TpG
	B3	T-C	563	283	TpC
Ibl-1260	F9	A-G	175	228	ApT
	F9	T-C	130	213	TpT
	A3	T-C	182	230	TpG
	H8	T-C	182	230	TpG
	C6	T-C	576	287	TpG
2630	H2	A-G	290	266	ApT
	E4	A-G	290	266	ApT
	G1	A-G	309	272	ApG
	G11	C-T	125	211	CpT
	G12	T-C	247	252	TpT
	B2	T-C	248	252	TpC
	B1	T-C	248	252	TpC
	E7	T-C	251	253	TpG
	F6	T-C	266	258	TpT
VACO425	G4	A-G	111	206	ApG
	F5	A-G	158	222	ApC
	F12	A-G	196	235	ApT
	A4	A-G	212	240	ApG
	A7	A-G	212	240	ApG
	E8	A-G	212	240	ApG
	H8	A-G	256	255	ApG
	F10	A-G	277	262	ApT
	C3	A-G	294	267	ApG
	D6	A-G	307	272	ApA
	D7	A-G	336	281	ApA
	E4	C-T	153	220	CpG
	G1	G-A	112	207	GpT
	F3	G-A	300	269	GpA
	C11	G-A	318	275	GpG
	D6	T-C	108	205	TpG
	D2	T-C	110	206	TpA
	H11	T-C	119	209	TpG
	G11	T-C	119	209	TpG
	B12	T-C	131	213	TpC
	H3	T-C	143	217	TpG
	C5	T-C	197	235	TpG
	G8	T-C	205	238	TpT
	D9	T-C	238	249	TpC
	D5	T-C	247	252	TpT
	B9	T-C	261	256	TpG
	A2	T-C	266	258	TpT

Cell Line	Position ID	APC Change	nt	Codon	Target Sequence
Ibl-1260	22	T-C	260	822	TpT
cont	50	T-C	295	834	TpC
	93	T-C	316	841	TpT
	39	T-C	324	843	TpT
	F3	T-C	325	844	TpC
	13	T-C	330	845	TpT
	19	T-C	387	864	TpC
	76	T-C	387	864	TpC
	91	T-C	414	873	TpT
	63	T-C	444	883	TpG
	84	T-C	455	887	TpT
	76	T-C	485	897	TpT
	16	T-C	588	931	TpT
	61	T-C	622	943	TpC
	B9	T-C	639	949	TpA
	50	T-C	645	950	TpT
	74	T-C	645	950	TpT
	32	T-C	693	966	TpT
	43	T-C	753	986	TpT
	95	T-C	756	987	TpG
	73	T-C	786	997	TpG
	2	T-C	813	1006	TpA
	24	T-C	831	1012	TpC
	2	T-C	879	1028	TpC
	37	T-C	888	1031	TpT
	37	T-C	909	1038	TpG
2630	B12	A-C	545	917	ApT
	C9	A-G	63	757	ApC
	E12	A-G	109	772	ApC
	B1	A-G	116	774	ApC
	E9	A-G	125	777	ApC
	E3	A-G	128	778	ApT
	H8	A-G	128	778	ApT
	F6	A-G	158	788	ApG
	E7	A-G	206	804	ApT
	F7	A-G	206	804	ApT
	F5	A-G	210	806	ApC
	E4	A-G	221	809	ApT
	G3	A-G	230	812	ApC
	E5	A-G	239	815	ApT
	B1	A-G	346	851	ApG
	F7	A-G	346	851	ApG
	E8	A-G	353	853	ApG
	B11	A-G	355	854	ApG
	B9	A-G	359	855	ApA
	F1	A-G	367	858	ApT
	E5	A-G	399	869	ApA
	G12	A-G	401	869	ApT
	C1	A-G	419	875	ApG
	H11	A-G	431	879	ApG
	H3	A-G	550	919	ApG
	F2	A-G	552	920	ApA
	F6	A-G	554	920	ApT
	A8	A-G	568	925	ApG
	G1	A-G	568	925	ApG
	E7	A-G	585	931	ApC
	A5	A-G	619	942	ApA
	C12	A-G	660	956	ApT
	F10	A-G	669	959	ApT
	G5	A-G	677	961	ApT
	D2	A-G	759	989	ApG
	G2	A-G	767	991	ApA
	F10	A-G	769	992	ApG
	G6	A-G	867	1025	ApA
	F7	A-G	927	1045	ApC
	G5	C-T	373	860	CpT
	G8	G-A	70	759	GpC
	F11	G-A	134	780	GpT
	D12	G-A	319	842	GpA
	G1	G-A	351	853	GpG
	D1	G-A	352	853	GpA
	F12	G-A	445	884	GpC
	D9	T-C	27	745	TpC
	A10	T-C	34	747	TpC
	H4	T-C	54	854	TpG
	A7	T-C	56	754	TpT
	B7	T-C	56	754	TpT
	E2	T-C	145	784	TpC
	H3	T-C	194	800	TpT
	A3	T-C	260	822	TpT
	B7	T-C	260	822	TpT
	A5	T-C	279	829	TpA
	B5	T-C	287	831	TpA
	D6	T-C	301	836	TpC
	F4	T-C	321	843	TpA

Cell Line	Position ID	APC Change	nt	Codon	Target Sequence
2630 cont	C5	T-C	330	846	TpT
	E8	T-C	374	860	TpA
	B10	T-C	382	863	TpA
	B11	T-C	414	874	TpT
	A2	T-C	476	894	TpG
	H6	T-C	476	894	TpG
	C3	T-C	571	926	TpC
	D2	T-C	582	930	TpA
	D12	T-C	582	930	TpA
	D6	T-C	604	937	TpT
	G12	T-C	673	960	TpC
	H5	T-C	754	987	TpC
	H9	T-C	783	997	TpT
	D1	T-C	836	1014	TpG
	H5	T-C	857	1021	TpA
	A1	T-C	873	1027	TpT
	B5	T-C	879	1029	TpC
	F5	T-C	879	1029	TpC
	E8	T-C	881	1029	TpT
	H2	T-C	907	1038	TpC
	B8	T-C	921	1043	TpC
	D11	T-G	30	746	TpG
	B8	T-G	374	860	TpA
VACO425	H1	A-C	447	885	ApG
	G9	A-G	45	751	ApT
	F2	A-G	58	755	ApG
	C6	A-G	63	757	ApC
	F10	A-G	102	770	ApT
	C4	A-G	139	782	ApA
	C2	A-G	140	782	ApG
	D12	A-G	163	790	ApG
	G5	A-G	167	791	ApC
	D2	A-G	228	812	ApG
	F12	A-G	264	824	ApC
	G8	A-G	303	837	ApT
	D11	A-G	309	839	ApG
	A11	A-G	312	840	ApA
	F9	A-G	339	849	ApG
	D7	A-G	346	851	ApG
	A7	A-G	357	855	ApG
	F1	A-G	390	866	ApG
	D11	A-G	394	867	ApC
	D12	A-G	401	869	ApT
	G9	A-G	433	880	ApT
	C11	A-G	447	885	ApG
	F5	A-G	473	893	ApA
	E9	A-G	480	896	ApG
	F4	A-G	501	903	ApG
	H3	A-G	501	903	ApG
	C9	A-G	505	904	ApG
	A7	A-G	548	918	ApG
	H1	A-G	564	924	ApA
	B1	A-G	567	925	ApA
	C12	A-G	659	955	ApA
	H11	A-G	669	959	ApT
	A2	A-G	723	977	ApG
	G11	A-G	734	980	ApA
	H1	A-G	764	990	ApT
	B10	A-G	768	992	ApA
	B4	A-G	798	1002	ApG
	D2	A-G	803	1003	ApC
	H8	A-G	803	1003	ApC
	C7	A-G	848	1018	ApT
	H5	A-G	896	1034	ApG
	H1	C-T	328	845	CpG
	F2	C-T	737	981	CpC
	A2	G-A	184	797	GpG
	G7	G-A	453	887	GpA
	A7	G-A	515	907	GpG
	A5	G-A	716	974	GpT
	H4	G-A	838	1015	GpA
	G7	T-A	183	797	TpG
	G12	T-C	48	752	TpC
	A8	T-C	50	752	TpT
	G6	T-C	103	770	TpC
	G10	T-C	103	770	TpC
	C11	T-C	111	773	TpT
	A2	T-C	255	821	TpG
	D5	T-C	279	829	TpA
	H9	T-C	321	843	TpA
	H4	T-C	330	846	TpT
	A1	T-C	331	846	TpC
	D10	T-C	372	860	TpC
	C9	T-C	434	880	TpC
	A10	T-C	485	897	TpT

Cell Line	Position ID	APC Change	nt	Codon	Target Sequence
VACO425 cont	E4	T-C	493	900	TpC
	G4	T-C	576	928	TpG
	A2	T-C	582	930	TpA
	G3	T-C	613	940	TpC
	A11	T-C	645	951	TpT
	H11	T-C	670	959	TpC
	G2	T-C	684	964	TpT
	H3	T-C	708	972	TpG
	E10	T-C	726	978	TpC
	E6	T-C	756	988	TpG
	C2	T-C	822	1010	TpA
	C12	T-C	876	1028	TpA
	B7	T-C	881	1029	TpT
	E8	T-C	924	1044	TpT

Appendix B

APC Mutations

Sample ID	Position ID	APC Change	nt	Codon	Target Sequence
2256N	C5	A-G	62	756	ApA
	A4	A-G	288	831	ApC
	A5	A-G	380	862	ApC
	H1	A-G	165	790	ApC
	E5	A-G	339	848	ApG
	E5	A-G	419	875	ApG
	E5	A-G	501	902	ApG
	G4	A-G	355	854	ApG
	C7	A-G	823	1010	ApG
	E7	A-G	913	1040	ApG
	A6	C-T	99	768	CpT
	H10	G-A	502	903	GpA
	G1	G-A	134	780	GpT
	A5	T-C	478	895	TpC
	B3	T-C	907	1038	TpC
	F12	T-C	748	985	TpC
	A8	T-C	753	986	TpT
2256T	A4	A-G	552	919	ApA
	B3	A-G	626	944	ApA
	E7	A-G	62	756	ApA
	E7	A-G	220	809	ApA
	F2	A-G	62	756	ApA
	F2	A-G	220	809	ApA
	B12	A-G	927	1044	ApC
	D12	A-G	631	946	ApC
	F11	A-G	503	903	ApC
	G5	A-G	503	903	ApC
	A11	A-G	735	980	ApC
	A2	A-G	355	854	ApG
	B2	A-G	667	958	ApG
	B4	A-G	462	889	ApG
	C3	A-G	78	761	ApG
	D3	A-G	355	854	ApG
	G2	A-G	339	848	ApG
	H5	A-G	355	854	ApG
	B2	A-G	223	810	ApG
	C1	A-G	471	892	ApG
	C3	A-G	896	1034	ApG
	A9	A-G	218	808	ApT
	C7	A-G	221	809	ApT
	E6	A-G	218	808	ApT
	E10	A-G	303	836	ApT
	F8	A-G	218	808	ApT
	F8	A-G	221	809	ApT
	G12	A-G	212	806	ApT
	H3	A-G	102	769	ApT
	H3	A-G	533	913	ApT
	A2	A-G	792	999	ApT
	A4	A-G	660	955	ApT
	A6	C-T	727	978	CpA
	E5	C-T	524	910	CpT
	B11	G-A	549	918	GpA
	H7	G-A	841	1016	GpA
	A1	G-A	841	1016	GpA
	A4	G-A	715	974	GpG
	B8	G-A	715	974	GpG
	C3	G-A	715	974	GpG
	C6	G-A	715	974	GpG
	C8	G-A	715	974	GpG
	C9	G-A	715	974	GpG
	B2	G-A	710	972	GpT
	E10	G-A	193	800	GpT
	C10	G-A	710	972	GpT
	B12	T-C	822	1009	TpA
	F10	T-C	382	863	TpA
	H4	T-C	519	908	TpA
	C4	T-C	571	926	TpC
	D7	T-C	325	844	TpC
	E10	T-C	295	834	TpC
	A11	T-C	907	1038	TpC
	A4	T-C	573	926	TpG
	B12	T-C	444	883	TpG
	C8	T-C	186	797	TpG
	B1	T-C	50	752	TpT
	C3	T-C	414	873	TpT
	E5	T-C	286	831	TpT
	G7	T-C	636	947	TpT

GPD1 Mutations

Sample ID	Position ID	GPD1 Change	nt	Codon	Target Sequence
2256N	C5	G-A	115	208	GpC
	D2	G-A	169	226	GpC
2256T	G4	A-G	290	266	ApT
	C10	C-T	229	246	CpC
	D1	C-T	303	270	CpC
	D9	C-T	203	237	CpC
	F2	C-T	203	237	CpC
	H6	C-T	284	264	CpC
	F7	C-T	326	278	CpG
	A4	C-T	214	241	CpT
	H3	C-T	142	217	CpT
	A5	G-A	136	215	GpA
	D2	G-A	198	235	GpA
	F4	G-A	109	206	GpT
	C5	T-C	248	252	TpC
	G2	T-C	215	241	TpC
	F4	T-C	270	259	TpG
	E5	T-C	143	217	TpG
	G6	T-C	148	219	TpT
2630N	A8	A-G	307	272	ApA
	D7	A-G	272	260	ApC
	B2	C-T	281	263	CpT
	G2	G-A	305	271	GpG
	B8	T-C	108	205	TpG
	E7	T-C	217	242	TpT
	F12	T-C	266	258	TpT
2630T	F7	A-C	294	267	ApG
	G8	C-T	279	262	CpA
	D6	C-T	167	225	CpG
	C1	G-A	268	259	GpC
	E12	T-C	206	238	TpC
	D12	T-C	138	215	TpG
	G11	T-A	225	244	TpG
	F2	T-C	322	277	TpT
2574N	G8	A-G	331	280	ApC
	E4	G-A	134	214	GpT
	E3	T-C	176	228	TpC
	F4	T-C	215	241	TpC
	F11	T-C	248	252	TpC
	E3	T-C	126	211	TpG
2574T	D1	G-A	339	282	ApG
	C12	G-A	226	245	GpG
2650N	B9	A-G	211	240	ApA
	B11	A-G	309	272	ApG
	E6	A-G	309	272	ApG
	E10	A-G	111	206	ApG
	E9	A-G	254	254	ApG
	H5	A-G	338	282	ApG
	H2	A-G	196	235	ApT
	C7	C-T	242	250	CpC
	D7	C-T	242	250	CpC
	B3	C-T	207	238	CpG
	F1	C-T	207	238	CpG
	F7	C-A	239	249	CpT
	G3	T-G	275	261	CpT
	H9	C-T	288	265	CpT
	D9	G-A	185	231	GpA
	G7	G-A	185	231	GpA
	D4	G-A	139	216	GpG
	F9	G-A	260	256	GpT
	E11	T-C	215	241	TpC
	D11	T-C	215	241	TpC
	F8	T-C	251	253	TpG
	G2	T-C	119	209	TpG
	A2	T-C	148	219	TpT
	G9	T-C	205	238	TpT
2650T	G4	A-G	137	215	ApT
	A7	G-A	140	216	GpC
	D2	G-A	192	233	GpG
	D10	G-A	127	212	GpG
	D10	G-A	265	258	GpT
	D11	G-A	112	207	GpT
	B9	T-G	110	206	TpA
	G3	T-C	286	265	TpG
	B2	T-C	148	219	TpT
	E9	T-C	130	213	TpT
	F6	T-C	205	238	TpT

Sample ID	Position ID	APC Change	nt	Codon	Target Sequence
2256T	E5	A-G	735	980	ApC
Continued	F4	A-G	230	812	ApC
	F10	A-G	735	980	ApC
	F8	A-G	792	999	ApT
	E1	C-T	727	978	CpA
	E5	C-T	524	910	CpT
	F2	G-A	340	849	GpA
	G4	G-A	841	1016	GpA
	E5	G-A	715	974	GpG
	F10	G-A	715	974	GpG
	G1	G-A	715	974	GpG
	E7	G-A	824	1010	GpT
	G10	G-A	710	972	GpT
	D2	T-C	879	1028	TpC
	H2	T-C	879	1028	TpC
	D9	T-C	762	989	TpG
	E6	T-C	775	994	TpT
2630N	B10	A-G	107	771	ApA
	B10	A-G	527	911	ApA
	E4	A-G	567	924	ApA
	F6	A-G	62	756	ApA
	C9	A-G	659	955	ApA
	C12	A-G	400	869	ApA
	D5	A-G	829	1012	ApA
	D8	A-G	767	991	ApA
	A9	A-G	593	933	ApC
	E1	A-G	277	828	ApC
	G1	A-G	585	930	ApC
	G12	A-G	98	768	ApC
	G1	A-G	599	935	ApC
	B10	A-G	501	902	ApG
	C9	A-G	548	918	ApG
	D4	A-G	313	840	ApG
	D4	A-G	343	850	ApG
	E2	A-G	419	875	ApG
	F10	A-G	105	770	ApG
	F11	A-G	105	770	ApG
	G11	A-G	105	770	ApG
	H1	A-G	899	1035	ApG
	E11	A-G	84	763	ApT
	B12	A-G	885	1030	ApT
	E5	A-G	887	1031	ApT
	E7	C-T	302	836	CpA
	E12	C-T	522	909	CpA
	E12	C-T	804	1003	CpC
	A9	C-T	33	746	CpT
	B10	G-A	499	902	GpA
	A6	G-C	91	766	GpC
	B3	G-A	808	1005	GpC
	B1	T-C	699	968	TpA
	F2	T-C	702	969	TpA
	C1	T-C	257	821	TpC
	F8	T-C	304	837	TpC
	G10	T-C	726	977	TpC
	G12	T-C	907	1038	TpC
	B4	T-C	183	796	TpG
	D11	T-C	255	820	TpG
	C4	T-C	111	772	TpT
	B8	T-C	645	950	TpT
	C1	T-C	711	972	TpT
	E9	T-C	684	963	TpT
2630T	C9	A-G	108	771	ApA
	C11	A-G	417	874	ApA
	B12	A-G	280	829	ApC
	A9	A-C	158	788	ApG
	B3	A-G	431	879	ApG
	E7	A-G	53	753	ApT
	C3	C-T	178	795	CpT
	C11	C-T	49	752	CpT
	D3	C-T	596	934	CpT
	D3	G-A	352	853	GpA
	E5	G-A	499	902	GpA
	A7	G-A	347	851	GpT
	C9	T-C	279	828	TpA
	A7	T-C	243	816	TpG
	C10	T-C	198	801	TpG
	E9	T-C	216	807	TpG

Sample ID	Position ID	GPD1 Change	nt	Codon	Target Sequence
2650T	C9	A-G	336	281	ApA
Continued	D9	G-A	185	231	GpA
	G9	G-A	185	231	GpA
	H6	G-A	226	245	GpG
	D9	T-C	110	206	TpA
	F10	T-C	133	214	TpG
	H10	T-C	251	253	TpG
1010N	D2	C-T	303	270	CpC
	E12	C-T	177	228	CpC
	G9	C-T	303	270	CpC
	B11	C-T	246	251	CpT
	D8	G-A	253	254	GpA
	F6	G-A	154	221	GpA
	D2	G-A	202	237	GpC
	E11	G-A	115	208	GpC
	F2	G-A	319	276	GpC
	C1	G-A	339	282	GpG
	E10	G-A	312	273	GpG
	F9	G-A	172	227	GpT
	G2	T-C	289	266	TpA
	B3	T-C	235	248	TpC
	E12	T-C	215	241	TpC
	G3	T-C	135	214	TpG
1010T	B11	A-G	163	224	ApA
	G11	C-T	162	223	CpA
	G12	C-T	156	221	CpA
	F7	C-T	228	245	CpC
	H3	C-G	228	245	CpC
	B8	C-T	178	229	CpG
	C2	G-A	316	275	GpA
	A4	G-A	120	209	GpG
	C8	G-A	145	218	GpG
	B9	G-A	263	257	GpT
	B5	T-C	131	213	TpC
2564N	G4	C-T	279	262	CpA
	C3	T-C	248	252	TpC
2564T	B6	C-T	273	260	CpC
	G4	C-T	167	225	CpG
	D3	C-T	187	232	CpT
	B11	G-A	253	254	GpA
	D6	G-A	185	231	GpA
	E12	G-A	174	227	GpA
	G10	G-A	306	271	GpA
	B11	G-A	202	237	GpC
	C6	G-A	227	245	GpC
	E12	G-A	268	259	GpC
	G10	G-A	146	218	GpC
	E9	G-A	183	230	GpG
	E9	G-A	265	258	GpT
	D6	T-C	188	232	TpC
1184T	D6	A-G	254	254	ApG
	G5	C-T	279	262	CpA
	H8	C-T	237	248	CpT
	E1	G-A	257	255	GpC
	H2	G-A	262	257	GpG
1962N	A3	A-G	111	206	ApG
	B7	A-G	333	280	ApG
	H11	G-A	306	271	GpA
	D1	G-A	335	281	GpC
	D2	T-C	205	238	TpT
1962T	D4	A-G	194	234	ApG
	C11	A-G	194	234	ApG
	H11	C-T	320	276	CpC
	H11	C-T	249	252	CpT
	E2	G-A	185	231	GpA
	F9	G-A	136	215	GpA
	D5	G-A	313	274	GpC
	G9	G-A	213	240	GpC
	G11	G-A	221	243	GpC
	A3	G-A	183	230	GpG
	D4	G-A	139	216	GpG
	C5	G-A	260	256	GpT
	D2	G-A	263	257	GpT
	E2	T-C	173	227	TpG
	G5	T-C	143	217	TpG
	G7	TDEL	324	277	TpG
	H9	T-C	311	273	TpG

Sample ID	Position ID	GPD1 Change	nt	Codon	Target Sequence
2630T	F12	A-G	338	848	ApA
Continued	G12	A-G	68	758	ApA
	D6	A-G	733	980	ApA
	D8	A-G	733	980	ApA
	G4	A-G	720	975	ApA
	A1	A-G	593	933	ApC
	A12	A-G	864	1023	ApC
	B11	A-G	864	1023	ApC
	F11	A-G	471	892	ApG
	G10	A-G	550	919	ApG
	H3	A-G	480	895	ApG
	H7	A-C	355	854	ApG
	G10	A-G	769	992	ApG
	G12	A-G	913	1040	ApG
	H6	A-G	929	1045	ApG
	A5	A-G	833	1013	ApT
	A12	A-G	689	965	ApT
	B11	A-G	689	965	ApT
	F12	A-G	845	1017	ApT
	F11	C-T	451	886	CpA
	E12	C-T	606	937	CpA
	E12	C-T	294	833	CpT
	B10	C-T	738	981	CpT
	G12	G-T	115	774	GpA
	G12	G-A	245	817	GpC
	F9	G-A	615	940	GpG
	H8	T-C	333	846	TpG
	B12	T-C	100	769	TpT
	G12	T-C	672	959	TpT
2395N	DB10	A-G	618	941	ApA
	DC5	A-G	867	1024	ApA
	DD5	A-G	733	980	ApA
	DD9	A-G	904	1037	ApA
	DF12	A-G	854	1020	ApA
	DA2	A-G	733	980	ApA
	DF5	A-G	854	1020	ApA
	BE1	A-G	68	758	ApA
	BF4	A-G	167	791	ApC
	BB9	A-G	109	772	ApC
	BE8	A-G	292	833	ApG
	BB7	A-G	95	767	ApG
	DA9	A-G	885	1030	ApT
	DB12	A-G	707	971	ApT
	DC2	A-G	680	962	ApT
	DC12	A-G	677	961	ApT
	DE3	A-G	640	949	ApT
	DF3	A-C	689	965	ApT
	DG12	A-G	842	1016	ApT
	DH11	A-G	677	961	ApT
	BF1	A-G	144	783	ApT
	BF9	A-G	267	824	ApT
	DF12	C-T	736	981	CpC
	DC2	G-A	757	988	GpA
	DC5	G-A	895	1034	GpA
	DF5	G-A	802	1003	GpA
	BA1	G-A	229	812	GpA
	DE11	G-A	642	949	GpC
	DA11	T-C	627	944	TpA
	DC2	T-C	699	968	TpA
	BD9	T-C	122	776	TpA
	BD9	T-C	190	799	TpA
	BF2	T-C	262	823	TpC
	BB9	T-C	194	800	TpT
2395T	DC5	A-G	653	953	ApA
	DC7	A-G	867	1024	ApA
	DF12	A-G	867	1024	ApA
	DA5	A-G	723	976	ApG
	DC2	A-G	858	1021	ApG
	DC5	A-G	723	976	ApG
	DE10	A-G	852	1019	ApG
	DE6	A-G	848	1018	ApT
	DF8	G-A	851	1019	GpA
	DF9	G-A	851	1019	GpA
	DB2	G-T	774	993	GpT
	DC2	G-T	774	993	GpT
	DC12	G-T	774	993	GpT
	DD5	G-T	774	993	GpT
	DF4	G-T	774	993	GpT

Sample ID	Position ID	GPD1 Change	nt	Codon	Target Sequence
1962T	F5	C-T	162	223	CpA
Continued	H11	C-T	326	278	CpG
1184T	D6	A-G	254	254	ApG
	G5	C-T	279	262	CpA
	H8	C-T	237	248	CpT
	E1	G-A	257	255	GpC
	H2	G-A	262	257	GpG
533N	C4	A-G	336	281	ApA
	D10	A-G	317	275	ApG
	E8	A-G	223	244	ApG
	G7	A-G	137	215	ApT
	B2	G-A	257	255	GpC
	C7	G-A	180	229	GpC
	F3	G-A	179	229	GpG
	D4	T-C	311	273	TpG
	F5	T-C	135	214	TpG
	D2	T-C	205	238	TpT
533T	B8	A-G	307	272	ApA
	B3	A-G	272	260	ApC
	H9	A-G	302	270	ApC
	A3	A-G	309	272	ApG
	C8	A-G	294	267	ApG
	H9	A-G	111	206	ApG
	C7	C-T	125	211	CpT
	D8	C-T	321	276	CpT
	A4	G-A	327	278	GpC
	A1	G-A	305	271	GpG
	B7	T-C	188	232	TpC
	B9	T-C	215	241	TpC
	C9	T-C	238	249	TpC
	G4	T-C	270	259	TpG
	F2	T-C	220	243	TpG
2494N	E7	A-G	573	360	ApG
	C6	C-T	153	220	CpG
	E10	C-T	216	241	CpT
	D10	G-A	241	250	GpC
	G4	G-A	180	229	GpC
	C12	G-A	598	369	GpC
	H3	T-C	176	228	TpC
	A3	T-C	231	246	TpG
	C6	T-C	150	219	TpG
	G5	T-C	148	219	TpT
2494T	H2	C-T	632	380	CpT
	A3	T-C	215	241	TpC
	E7	T-C	248	252	TpC
	G12	T-C	645	384	TpC
	D10	T-C	225	244	TpG
2486N	F2	A-G	636	381	ApC
	F9	A-G	223	244	ApG
	E4	C-T	650	386	CpA
	G11	C-T	667	392	CpA
	E6	C-T	298	269	CpG
	F11	G-A	166	225	GpC
	H5	G-A	168	225	GpG
	B1	T-C	250	253	TpT
2486T	D5	C-A	288	265	CpT
2491N	G7	A-G	666	391	ApC
	C3	A-G	254	254	ApG
	E11	A-G	290	266	ApT
	F8	C-T	298	269	CpG
	A6	C-T	230	246	CpT
	F12	G-A	193	234	GpA
	B4	G-A	139	216	GpG
	F8	T-C	238	249	TpC
	B12	T-C	311	273	TpG
	D11	T-C	291	266	TpG
	D7	T-C	266	258	TpT
2490N	A10	A-G	317	275	ApG
	C3	A-G	201	236	ApG
	B3	C-T	616	375	CpG
	A10	G-A	185	231	GpA
	A11	G-A	169	226	GpC
	H3	T-C	215	241	TpC
	H9	T-C	235	248	TpC
	C5	T-C	275	261	TpG
	C11	T-C	264	257	TpG
	E10	T-C	225	244	TpG
	G6	T-C	275	261	TpG

Sample ID	Position ID	GPD1 Change	nt	Codon	Target Sequence
2395T	DH6	A-G	680	962	ApT
Continued	DG7	G-T	774	993	GpT
	DH10	G-T	774	993	GpT
	DB10	T-C	889	1032	TpC
	DG12	T-C	881	1029	TpT
2574N	D8	A-G	65	757	ApA
	F11	A-G	447	884	ApG
	B3	A-G	543	916	ApG
	C9	A-G	858	1021	ApG
	F12	A-G	554	920	ApT
	B7	A-G	792	999	ApT
	F1	A-G	554	920	ApT
	G6	A-G	842	1016	ApT
	G8	A-G	401	869	ApT
	A9	C-T	211	806	CpA
	D2	C-T	148	785	CpA
	F8	C-T	810	1005	CpC
	D1	C-T	180	795	CpT
	D11	C-T	180	795	CpT
	B11	G-A	124	777	GpA
	H9	G-A	124	777	GpA
	A3	G-A	171	792	GpC
	C8	G-A	283	830	GpT
	G12	G-A	477	894	GpT
	C9	T-C	717	974	TpA
	D7	T-C	476	894	TpG
2574T	BH5	A-G	118	775	ApA
	DD12	A-G	626	944	ApA
	BD6	A-G	125	777	ApC
	BA11	A-G	280	829	ApC
	DF5	A-G	803	1003	ApC
	BD1	A-G	170	792	ApG
	DA8	A-G	899	1035	ApG
	DE8	A-G	913	1040	ApG
	DG12	A-G	721	976	ApG
	DH2	A-T	667	958	ApG
	DH10	A-G	721	976	ApG
	CA10	A-G	562	923	ApG
	CF1	A-G	431	879	ApG
	CF10	A-G	322	843	ApG
	CF12	A-G	505	904	ApG
	BA7	A-G	89	765	ApT
	DD8	A-G	654	953	ApT
	DC2	A-G	792	999	ApT
	DF2	A-G	792	999	ApT
	CA5	A-G	484	897	ApT
	BC7	C-T	246	817	CpA
	BA10	C-T	99	768	CpT
	BH8	C-T	92	766	CpT
	DD11	G-T	732	979	GpA
	DH7	G-A	895	1034	GpA
	BE8	G-A	38	748	GpC
	DG3	G-A	629	945	GpG
	BA7	G-A	152	786	GpT
	BC10	T-C	101	769	TpA
	BG5	T-C	147	784	TpC
	BH2	T-C	93	766	TpC
	DA7	T-C	879	1028	TpC
	DA9	T-C	879	1028	TpC
	DC6	T-C	879	1028	TpC
	DD6	T-C	879	1028	TpC
	BC12	T-C	41	749	TpG
	DF12	T-C	648	951	TpG
	CF12	T-C	537	914	TpG
	CH9	T-C	428	878	TpG
	BD10	T-C	195	800	TpT
	BH3	T-C	130	779	TpT
	BH9	T-C	194	800	TpT
	BD11	T-C	300	835	TpT
	DC5	T-C	888	1031	TpT
	DE11	T-C	775	994	TpT
	CA5	T-C	510	905	TpT
	CE6	T-C	414	873	TpT
2650N	BB12	C-T	136	781	CpC
	BB5	C-T	99	768	CpT
	BA11	G-A	91	766	GpC
	BA5	T-C	57	754	TpA
	BA11	T-C	90	765	TpG

Sample ID	Position ID	GPD1 Change	nt	Codon	Target Sequence
2490N	H6	G-A	617	375	GpA
Continued	G4	T-C	576	361	TpG

Sample ID	Position ID	GPD1 Change	nt	Codon	Target Sequence
2650N	BD7	A-G	312	839	ApA
Continued	CE12	A-G	345	850	ApA
	DB3	A-G	772	993	ApA
	DB8	A-G	854	1020	ApA
	DG10	A-G	768	991	ApA
	DG10	A-G	719	975	ApA
	DB7	A-G	854	1020	ApA
	BC8	A-G	98	768	ApC
	BD7	A-G	165	790	ApC
	CB6	A-G	583	930	ApC
	CE6	A-G	383	863	ApC
	CF6	A-G	490	899	ApC
	DD10	A-G	803	1003	ApC
	DE11	A-G	803	1003	ApC
	DF4	A-G	803	1003	ApC
	BE10	A-G	163	790	ApG
	BD4	A-G	154	787	ApG
	BD9	A-G	307	838	ApG
	CA2	A-G	501	902	ApG
	CB9	A-G	390	865	ApG
	CE11	A-G	390	865	ApG
	DB3	A-G	721	976	ApG
	DE8	A-G	691	966	ApG
	DE10	A-G	313	840	ApG
	DF7	A-G	703	970	ApG
	DB1	A-G	660	955	ApT
	DB12	A-G	602	936	ApT
	DD2	A-G	669	958	ApT
	DE2	A-G	821	1009	ApT
	DF1	A-G	875	1027	ApT
	DF9	A-G	792	999	ApT
	DF9	A-G	842	1016	ApT
	DG8	A-C	792	999	ApT
	DB5	A-G	835	1014	ApT
	DC3	A-G	640	949	ApT
	DC4	A-G	932	1046	ApT
	BC8	C-T	166	791	CpA
	CC6	C-T	395	867	CpA
	CH10	C-T	395	867	CpA
	DB2	C-T	800	1002	CpC
	BF10	C-T	294	833	CpT
	DA7	C-T	750	985	CpT
	DE2	C-T	750	985	CpT
	BD2	G-A	273	826	GpA
	BG6	G-A	124	777	GpA
	DC11	G-A	859	1022	GpA
	CB6	G-A	481	896	GpC
	DE10	G-A	914	1040	GpG
	DF6	G-A	914	1040	GpG
	DD10	G-A	694	967	GpT
	BD3	T-C	181	796	TpA
	BF11	T-C	101	769	TpA
	BE8	T-C	321	842	TpA
	DG9	A-G	690	965	TpA
	DG10	T-C	834	1013	TpA
	CF3	T-C	517	908	TpC
	DE2	T-C	748	985	TpC
	DF3	T-C	889	1032	TpC
	DG2	T-C	613	940	TpC
	DH1	T-C	622	943	TpC
	BC8	T-G	243	816	TpG
	BG5	T-C	90	765	TpG
	CA1	T-C	576	927	TpG
	CA2	T-C	576	927	TpG
	CE8	T-C	539	915	TpG
	CE9	T-C	573	926	TpG
	CG7	T-C	555	920	TpG
	CH8	T-C	476	894	TpG
	DE1	T-C	648	951	TpG
	DE12	T-C	705	970	TpG
	DF5	T-C	648	951	TpG
	DF7	T-C	714	973	TpG
	BA5	T-C	234	813	TpT
	CB9	T-C	414	873	TpT
	CD7	T-C	411	872	TpT
	CE11	T-C	414	873	TpT
	DH8	T-C	743	983	TpT
	DH10	T-C	743	983	TpT

Sample ID	Position ID	GPD1 Change	nt	Codon	Target Sequence
2650N Continued	DE8	A-G	733	980	ApA
	DF5	A-G	803	1003	ApC
	DF12	A-G	794	1000	ApC
	DD9	A-G	660	955	ApT
	DF9	G-A	914	1040	GpG
	DD9	T-C	634	947	TpG
	DC10	T-C	901	1036	TpT
	DG7	T-C	881	1029	TpT
2650T	CF12	A-G	552	919	ApA
	DA1	A-G	720	975	ApA
	DB2	A-G	734	980	ApA
	DB6	A-G	676	961	ApA
	DE8	A-G	688	965	ApA
	BE4	A-G	165	790	ApC
	BE11	A-G	288	831	ApC
	CF9	A-G	541	916	ApC
	CH2	A-G	541	916	ApC
	DB6	A-G	864	1023	ApC
	BE1	A-G	170	792	ApG
	BH4	A-G	170	792	ApG
	CA4	A-G	562	923	ApG
	CA8	A-G	452	886	ApG
	CF12	A-G	396	867	ApG
	CH4	A-G	462	889	ApG
	CH9	A-G	505	904	ApG
	DG11	A-G	769	992	ApG
	DE2	A-G	759	988	ApG
	DE5	A-G	723	976	ApG
	BD10	AT-GC	269/270	825	ApT
	BB7	A-G	239	815	ApT
	CC3	A-G	341	849	ApT
	DB7	A-G	812	1006	ApT
	DD6	A-G	654	953	ApT
	DF4	A-G	633	946	ApT
	DG12	A-G	640	949	ApT
	DG12	A-G	761	989	ApT
	DB5	A-G	817	1008	ApT
	DB5	A-G	885	1030	ApT
	DG11	A-G	640	949	ApT
	BD10	C-A	204	803	CpA
	BG3	C-T	263	823	CpA
	BC9	C-T	265	824	CpC
	CH3	C-T	450	885	CpC
	CG8	C-T	494	900	CpT
	BB8	G-A	115	774	GpA
	CG11	G-A	356	854	GpA
	DG12	G-A	897	1034	GpC
	BD1	G-A	256	821	GpT
	DG5	G-A	698	968	GpT
	CA2	T-C	519	908	TpA
	CC6	T-C	382	863	TpA
	DA1	T-C	639	948	TpA
	DA6	T-C	886	1031	TpA
	DC11	T-C	857	1021	TpA
	DF6	T-C	712	973	TpA
	DG2	T-C	699	968	TpA
	BF3	T-C	147	784	TpC
	CC9	T-C	434	880	TpC
	CD7	T-C	327	844	TpC
	CF5	T-C	517	908	TpC
	CF11	T-C	434	880	TpC
	DB3	T-C	695	967	TpC
	DE1	T-C	748	985	TpC
	DE8	T-C	921	1042	TpC
	DH8	T-C	726	977	TpC
	DA6	T-C	889	1032	TpC
	DD7	T-C	925	1044	TpC
	BA4	T-C	213	806	TpG
	CA6	T-C	467	891	TpG
	CB4	T-C	573	926	TpG
	DB8	T-C	762	989	TpG
	DC1	T-C	825	1010	TpG
	DE1	T-C	693	966	TpG
	DF8	T-C	744	983	TpG
	BF8	T-C	261	822	TpT
	CD9	T-C	348	851	TpT
	DA8	T-C	783	996	TpT
	DC6	T-C	753	986	TpT

Sample ID	Position ID	GPD1 Change	nt	Codon	Target Sequence
2650T Cont.	DH1	CINS	738-739	981	CpT
	DD8	T-C	603	936	TpT
1010N	BF9	A-G	127	778	ApA
	CA5	A-G	418	875	ApA
	CA6	A-G	400	869	ApA
	CA10	A-G	418	875	ApA
	CC2	A-G	400	869	ApA
	DD6	A-G	719	975	ApA
	DE6	A-G	664	957	ApA
	BA1	A-G	230	812	ApC
	BB6	A-G	63	756	ApC
	BD3	A-G	165	790	ApC
	DH4	A-G	862	1023	ApC
	BB2	A-G	140	782	ApG
	BC4	A-G	307	838	ApG
	BG3	A-G	78	761	ApG
	BH9	A-G	313	840	ApG
	BC2	A-G	163	790	ApG
	BD4	A-G	69	758	ApG
	DB7	A-G	919	1042	ApG
	DE6	A-G	691	966	ApG
	DE8	A-G	852	1019	ApG
	DG5	A-G	852	1019	ApG
	BA9	A-G	267	824	ApT
	BF9	A-G	89	765	ApT
	BH9	A-G	250	819	ApT
	DA7	A-G	689	965	ApT
	DD12	A-G	730	979	ApT
	BA2	C-T	201	802	CpA
	BF11	C-T	305	837	CpA
	BH1	C-T	138	781	CpA
	BH1	C-T	43	750	CpC
	BE12	C-T	137	781	CpC
	CA9	C-T	578	928	CpC
	BA8	C-T	33	746	CpT
	BF2	C-T	73	760	CpT
	BH2	G-A	214	807	GpA
	BB8	G-A	209	805	GpA
	CA10	G-A	469	892	GpA
	CA10	G-A	469	892	GpA
	CC5	G-A	365	857	GpA
	DA4	G-A	838	1015	GpA
	DA6	G-A	741	982	GpA
	BB6	G-A	32	746	GpC
	BG3	G-A	314	840	GpC
	BC2	G-A	244	817	GpG
	BF1	G-A	141	782	GpG
	DA5	G-A	629	945	GpG
	DB9	G-A	724	977	GpG
	DG5	G-A	715	974	GpG
	BF9	G-A	225	810	GpT
	DB2	G-A	694	967	GpT
	DE11	G-A	694	967	GpT
	DH10	G-A	710	972	GpT
	DH10	G-T	774	993	GpT
	BC3	T-C	57	754	TpA
	DF11	T-C	751	986	TpA
	DH2	T-C	712	973	TpA
	BA8	T-C	226	811	TpC
	BG1	T-C	295	834	TpC
	DE7	T-C	637	948	TpC
	DE8	T-C	673	960	TpC
	CB12	T-C	539	915	TpG
	DB4	T-C	693	966	TpG
	DE6	T-C	648	951	TpG
	DF2	T-C	641	949	TpG
	DH6	T-C	756	987	TpG
	BC9	T-C	300	835	TpT
	DA10	T-C	645	950	TpT
	DH2	T-C	603	936	TpT
1010T	BD5	A-G	139	782	ApA
	BB2	A-G	264	823	ApC
	BC5	A-G	288	831	ApC
	BD2	A-G	133	780	ApG
	BA11	A-G	119	775	ApT
	BA5	C-T	43	750	CpC
	BD5	C-T	297	834	CpT
	BH5	G-A	319	842	GpA

Sample ID	Position ID	GPD1 Change	nt	Codon	Target Sequence
1010T	CH9	A-G	553	920	ApA
Continued	DA8	A-G	931	1046	ApA
	DB7	A-G	729	978	ApA
	DB12	A-G	871	1026	ApA
	DC2	A-G	733	980	ApA
	DC10	A-G	617	941	ApA
	DE1	A-G	624	943	ApA
	DG10	A-C	653	953	ApA
	DG10	A-G	688	965	ApA
	DB2	A-G	734	980	ApA
	CE5	A-G	568	925	ApG
	CG2	A-G	353	853	ApG
	CG4	A-G	447	884	ApG
	CG12	A-T	452	886	ApG
	DC10	A-G	697	968	ApG
	DE4	A-G	665	957	ApG
	DE12	A-G	877	1028	ApG
	DF12	A-G	665	957	ApG
	DG9	A-G	759	988	ApG
	DC9	A-G	773	993	ApG
	DC7	A-G	669	958	ApT
	DE9	A-G	640	949	ApT
	DG9	A-G	689	965	ApT
	CE3	C-T	580	929	CpA
	CF9	C-T	590	932	CpA
	DD5	C-T	606	937	CpA
	CE3	C-T	443	883	CpT
	DE2	C-T	755	987	CpT
	DH8	C-T	110	772	CpT
	CD12	G-A	354	853	GpA
	DF4	G-A	679	962	GpA
	CD2	G-A	448	885	GpC
	CD5	G-A	391	866	GpC
	CG2	G-A	448	885	GpC
	CD6	G-A	538	915	GpT
	DA2	G-A	694	967	GpT
	DB12	G-A	725	977	GpT
	BG5	T-C	48	751	TpC
	CH5	T-C	412	873	TpC
	CH9	T-C	327	844	TpC
	DB4	T-C	925	1044	TpC
	CE1	T-C	573	926	TpG
	CD10	T-C	368	858	TpT
	CE6	T-C	324	843	TpT
	CG2	T-C	560	922	TpT
	DG11	T-C	711	972	TpT
1299N	BB3	A-G	108	771	ApA
	BB6	A-G	157	788	ApA
	BB8	A-T	77	761	ApA
	BB9	A-G	108	771	ApA
	DA6	A-G	829	1012	ApA
	DB5	A-G	599	935	ApC
	BB6	A-G	69	758	ApG
	BB4	C-T	94	767	CpA
	BD3	C-T	94	767	CpA
	DA6	C-T	898	1035	CpA
	BB4	C-T	137	781	CpC
	BD1	C-T	110	772	CpT
	BD2	C-T	99	768	CpT
	BD3	C-T	99	768	CpT
	BD1	G-A	60	755	GpA
	DB8	G-A	630	945	GpA
	DC10	G-A	630	945	GpA
	DA3	G-A	782	996	GpT
	DC1	G-A	698	968	GpT
	DC11	G-A	716	974	GpT
	DA6	T-C	598	935	TpA
	DB8	T-C	879	1028	TpC
	DC4	T-C	879	1028	TpC
	DC10	T-C	879	1028	TpC
	DA6	T-C	705	970	TpG
	DC6	T-C	783	996	TpT
	DB2	T-C	672	959	TpT
	DB2	T-C	645	950	TpT
1299T	BF3	A-G	264	823	ApC
	BF6	A-G	264	823	ApC
	BE5	A-G	215	807	ApT
	BE1	G-A	293	833	GpC

Sample ID	Position ID	GPD1 Change	nt	Codon	Target Sequence
1299T	BG7	A-G	139	782	ApA
Continued	DF4	A-G	854	1020	ApA
	DF5	A-G	619	942	ApA
	DG3	A-G	854	1020	ApA
	BG3	A-G	69	758	ApG
	BG4	A-G	228	811	ApG
	DE7	A-G	703	970	ApG
	DF8	A-G	703	970	ApG
	DG3	A-G	852	1019	ApG
	DE2	A-G	860	1022	ApT
	DE8	A-G	669	958	ApT
	DF4	A-G	932	1046	ApT
	DG3	A-G	932	1046	ApT
	DH6	A-G	860	1022	ApT
	DG7	A-G	835	1014	ApT
	BG4	C-T	281	829	CpA
	BG2	C-G	305	837	CpA
	DF11	C-T	866	1024	CpA
	DG6	C-T	866	1024	CpA
	DG5	C-G	650	952	CpC
	DF1	C-T	856	1021	CpT
	DF2	C-T	856	1021	CpT
	DG11	C-T	856	1021	CpT
	BG7	G-A	187	798	GpA
	DE8	T-C	861	1022	TpA
	BE1	T-C	179	795	TpC
	BG7	T-C	135	780	TpC
	DG2	T-C	695	967	TpC
	DF3	T-C	756	987	TpG
	DG10	T-C	756	987	TpG
2564N	C9	A-G	400	869	ApA
	E7	A-T	345	850	ApA
	F5	A-G	400	869	ApA
	A2	A-G	439	882	ApC
	A8	A-G	490	899	ApC
	B5	A-G	583	930	ApC
	C6	A-G	523	910	ApC
	E2	A-G	116	774	ApC
	D5	A-G	123	776	ApG
	D9	A-G	396	867	ApG
	E5	A-G	357	854	ApG
	H10	A-G	628	945	ApG
	H6	A-G	554	920	ApT
	D12	C-T	866	1024	CpA
	F1	C-T	137	781	CpC
	G4	C-T	203	803	CpC
	H3	C-T	71	759	CpC
	G10	C-T	361	856	CpG
	D2	G-A	365	857	GpA
	F8	G-A	314	840	GpC
	C8	G-A	709	972	GpG
	B7	G-A	477	894	GpT
	B8	G-A	536	914	GpT
	G3	G-A	329	845	GpT
	F3	G-A	770	992	GpT
	D6	T-C	51	752	TpC
	F7	T-C	789	998	TpC
	H8	T-C	613	940	TpC
	C3	T-C	535	914	TpG
	B2	T-C	714	973	TpG
	B1	T-C	20	742	TpT
	F8	T-C	50	752	TpT
	F10	T-C	300	835	TpT
	H9	T-C	510	905	TpT
	G4	T-C	645	950	TpT
	H4	T-C	775	994	TpT
2564T	DA3	A-G	883	1030	ApA
	DA7	A-G	798	1001	ApG
	DA11	A-G	759	988	ApG
	DB2	A-G	769	992	ApG
	DB4	A-G	628	945	ApG
	DC2	A-G	697	968	ApG
	DD3	A-G	700	969	ApG
	DD10	A-G	798	1001	ApG
	DE1	A-G	703	970	ApG
	DB10	C-T	865	1024	CpC
	DC11	G-A	716	974	GpT
	DD9	T-C	909	1038	TpG

Sample ID	Position ID	GPD1 Change	nt	Codon	Target Sequence
2564T	DF7	A-G	828	1011	ApA
Continued	DH4	A-G	618	941	ApA
	BB8	A-G	67	758	ApA
	CB6	A-C	470	892	ApA
	CD2	A-G	473	893	ApA
	DH3	A-G	552	919	ApA
	BD1	A-G	200	802	ApC
	DG1	A-G	769	992	ApG
	DG6	A-G	769	992	ApG
	DG7	A-G	759	988	ApG
	DA2	A-G	759	988	ApG
	DD3	A-G	700	969	ApG
	BB4	A-G	75	760	ApG
	BA1	A-G	170	792	ApG
	DC10	A-G	419	875	ApG
	DH3	A-G	309	838	ApG
	DH9	A-G	501	902	ApG
	DG7	A-C	689	965	ApT
	DA2	A-C	689	965	ApT
	DH1	C-T	898	1035	CpA
	BF4	C-T	302	836	CpA
	BG8	C-T	231	812	CpA
	CD8	C-T	522	909	CpA
	DC9	C-T	522	909	CpA
	CA3	C-T	578	928	CpC
	CB5	C-T	578	928	CpC
	BA5	G-A	319	842	GpA
	BA11	G-A	164	790	GpA
	BB4	G-T	344	850	GpA
	BE8	G-C	106	771	GpA
	CB4	GADEL	358-359	855	GpA
	DH4	G-A	788	998	GpT
	CE2	G-A	329	845	GpT
	DH1	T-C	598	935	TpA
	DE3	T-C	909	1038	TpG
	DE3	T-C	909	1038	TpG
	BG1	T-C	251	819	TpG
	BG7	T-C	183	796	TpG
	DD7	T-C	888	1031	TpT
	DH6	T-C	888	1031	TpT
	BF5	T-C	111	772	TpT
	BH7	T-C	236	814	TpT
	CE4	T-C	597	934	TpT
1184N	CE6	A-G	379	862	ApA
	DA1	A-G	617	941	ApA
	DD4	A-G	844	1017	ApA
	DF4	A-G	768	991	ApA
	DB1	A-G	855	1020	ApC
	DF4	A-G	864	1023	ApC
	BE6	A-G	163	790	ApG
	BH4	A-G	78	761	ApG
	CC2	A-G	355	854	ApG
	CC8	A-G	497	901	ApG
	CE3	A-G	423	876	ApG
	DB4	A-G	773	993	ApG
	DC9	A-G	852	1019	ApG
	DF2	A-G	858	1021	ApG
	DF11	A-G	665	957	ApG
	DA8	A-G	872	1026	ApT
	DC8	A-G	669	958	ApT
	DE4	A-G	689	965	ApT
	DE12	A-G	932	1046	ApT
	BA1	C-T	281	829	CpA
	DC2	C-T	696	967	CpA
	DF9	C-T	727	978	CpA
	DB7	C-T	638	948	CpT
	DE2	C-T	638	948	CpT
	CE4	G-A	556	921	GpC
	CC11	G-A	514	907	GpG
	CF10	G-A	514	907	GpG
	CA3	T-C	374	860	TpA
	DE5	T-C	806	1004	TpA
	DE10	T-C	717	974	TpA
	DD7	T-C	921	1042	TpC
	BA1	T-C	186	797	TpG
	CC9	T-C	455	887	TpT
	CG7	T-C	455	887	TpT
	DD4	T-C	655	954	TpT

Sample ID	Position ID	GPD1 Change	nt	Codon	Target Sequence
1184N Continued	DE12	T-C	901	1036	TpT
	DG6	G-A	629	945	GpG
	DB6	T-C	925	1044	TpC
	DD4	T-C	695	967	TpC
1184T	BA11	A-G	174	793	ApA
	BE4	A-G	174	793	ApA
	DC1	A-G	625	944	ApA
	DC1	A-G	728	978	ApA
	DD4	A-G	626	944	ApA
	DE12	A-G	854	1020	ApA
	DE12	A-G	918	1041	ApA
	BA5	A-G	264	823	ApC
	BA11	A-G	280	829	ApC
	BE4	A-G	280	829	ApC
	DC12	A-T	631	946	ApC
	BD2	A-G	123	776	ApG
	BD9	A-G	175	794	ApG
	BE8	A-G	140	782	ApG
	CB4	A-G	550	919	ApG
	CD12	A-G	565	924	ApG
	CG4	A-G	452	886	ApG
	CG4	A-G	497	901	ApG
	CH10	A-G	505	904	ApG
	DD1	A-G	665	957	ApG
	DD2	A-G	852	1019	ApG
	DE12	A-G	929	1045	ApG
	DG5	A-G	781	996	ApG
	DD8	A-G	835	1014	ApT
	DE3	A-G	835	1014	ApT
	DG5	A-G	677	961	ApT
	CE4	C-T	479	895	CpA
	CE4	C-T	496	901	CpA
	DH4	C-T	632	946	CpA
	DF8	C-T	650	952	CpC
	DA11	G-A	851	1019	GpA
	DB1	G-A	679	962	GpA
	CD12	G-A	391	866	GpC
	CD3	G-A	468	891	GpG
	CG4	G-A	364	857	GpG
	DD11	T-C	656	954	TpA
	DE10	T-C	639	948	TpA
	CH1	T-G	495	900	TpC
	DB5	T-C	637	948	TpC
	DC10	T-C	726	977	TpC
	DG1	T-C	726	977	TpC
	DG9	T-C	879	1028	TpC
	CG4	T-C	576	927	TpG
	DC5	T-C	744	983	TpG
	DH1	T-A	778	995	TpG
	BE1	T-C	100	769	TpT
	CB4	T-C	485	897	TpT
	CD12	T-C	427	878	TpT
	DD1	T-C	645	950	TpT
	DF9	T-C	888	1031	TpT
	DH8	T-C	597	934	TpT
3603N	BD7	A-G	169	792	ApA
	CA3	A-G	564	923	ApA
	BC4	A-G	264	823	ApC
	BC8	A-G	288	831	ApC
	BC8	A-G	264	823	ApC
	BD3	A-G	200	802	ApC
	BE6	A-G	253	820	ApC
	BC8	A-G	264	823	ApC
	BC1	A-G	163	790	ApG
	BD7	A-G	69	758	ApG
	BD12	A-G	154	787	ApG
	CB1	A-G	353	853	ApG
	BF5	A-G	218	808	ApT
	BA8	C-T	203	803	CpC
	BD1	C-T	29	745	CpT
	CD3	G-A	340	849	GpA
	BA5	G-A	42	749	GpC
	CB3	G-A	457	888	GpC
	CD2	G-A	370	859	GpG
	CB10	T-C	582	929	TpA
	BH9	T-C	226	811	TpC
	CA10	T-C	464	890	TpC
	CA1	T-C	476	894	TpG

Sample ID	Position ID	GPD1 Change	nt	Codon	Target Sequence
3603N	CB6	A-G	553	920	ApA
Continued	DB3	A-G	733	980	ApA
	DE12	A-G	676	961	ApA
	CG7	A-G	523	910	ApC
	DA8	A-G	735	980	ApC
	CB6	A-G	547	918	ApG
	DA9	A-G	773	993	ApG
	DC9	A-G	723	976	ApG
	DG10	A-G	700	969	ApG
	DE2	A-G	835	1014	ApT
	DF5	A-G	792	999	ApT
	DH7	A-G	932	1046	ApT
	CE2	C-T	465	890	CpA
	DE7	C-T	790	999	CpA
	DG9	C-T	922	1043	CpC
	DC12	C-T	755	987	CpT
	DB6	C-T	908	1038	CpT
	CG8	G-A	352	853	GpA
	DB10	G-A	838	1015	GpA
	DB5	G-A	903	1036	GpA
	DE9	G-A	895	1034	GpA
	CD8	G-A	362	856	GpC
	DE9	G-A	779	995	GpC
	CE4	G-A	370	859	GpG
	DD12	G-A	715	974	GpG
	CH6	T-C	519	908	TpA
	CD3	T-C	372	859	TpC
	CB6	T-C	571	926	TpC
	DA7	T-C	613	940	TpC
	DD12	T-C	754	987	TpC
	DE5	T-C	849	1018	TpG
	CC1	T-C	597	934	TpT
	DC6	T-C	924	1043	TpT
	DG5	T-C	603	936	TpT
3603T	CF10	A-G	338	848	ApA
	CH6	A-G	417	874	ApA
	DD3	G-A	917	1041	ApA
	DD8	A-G	746	984	ApA
	CC9	A-G	541	916	ApC
	CE5	A-G	360	855	ApC
	CH2	A-G	360	855	ApC
	CH6	A-G	439	882	ApC
	DD8	A-G	631	946	ApC
	DF7	A-G	927	1044	ApC
	CC12	A-G	419	875	ApG
	CF10	A-G	550	919	ApG
	DC6	A-G	665	957	ApG
	DE2	A-G	721	976	ApG
	DF3	A-G	723	976	ApG
	CA9	A-G	587	931	ApT
	DA7	A-C	860	1022	ApT
	DA9	A-G	713	973	ApT
	DC2	A-G	860	1022	ApT
	DD12	A-G	792	999	ApT
	DG6	A-G	660	955	ApT
	DG8	A-G	707	971	ApT
	DF8	C-T	810	1005	CpC
	CC9	C-T	326	844	CpT
	CC6	G-A	566	924	GpA
	CD10	G-A	422	876	GpA
	DC9	G-A	642	949	GpC
	CF11	G-A	498	901	GpG
	DB9	G-A	782	996	GpT
	DD12	G-A	878	1028	GpT
	CC11	T-C	372	859	TpC
	DH8	T-C	748	985	TpC
	CE3	T-C	539	915	TpG
	CC6	T-C	411	872	TpT
	DE5	T-C	645	950	TpT
1962N	BB5	A-G	170	792	ApG
	BB10	A-G	318	841	ApG
	CA9	A-G	353	853	ApG
	CA10	A-G	346	851	ApG
	CA5	A-G	367	858	ApT
	BA12	C-T	265	824	CpC
	BB9	T-C	321	842	TpA
	CA1	T-C	374	860	TpA
	BB7	T-C	226	811	TpC

Sample ID	Position ID	GPD1 Change	nt	Codon	Target Sequence
1962N Continued	DG10	A-G	768	991	ApA
	CC8	A-G	405	870	ApG
	DG4	A-G	665	957	ApG
	CB5	G-A	515	907	GpG
	DG5	T-C	748	985	TpC
	BB7	T-C	316	841	TpT
	CC10	T-C	414	873	TpT
1962T	BF9	A-T	169	792	ApA
	CG3	A-G	500	902	ApA
	DB10	A-G	734	980	ApA
	DD10	A-G	768	991	ApA
	CG8	A-G	380	862	ApC
	DC4	A-G	662	956	ApC
	BF10	A-G	292	833	ApG
	BH1	A-G	154	787	ApG
	BE9	C-T	64	757	CpA
	DA9	C-T	696	967	CpA
	DA12	C-T	827	1011	CpA
	DB9	C-T	827	1011	CpA
	DC3	C-T	916	1041	CpA
	BH8	C-T	137	781	CpC
	CG2	C-T	413	873	CpT
	DC8	G-A	808	1005	GpC
	BG7	G-A	244	817	GpG
	DD8	G-A	715	974	GpG
	BF3	G-A	185	797	GpT
	DB11	G-A	774	993	GpT
	DD8	G-A	701	969	GpT
	DE8	G-A	635	947	GpT
	BH1	T-C	219	808	TpA
	DC2	T-C	784	997	TpA
	DD6	T-C	857	1021	TpA
	DE4	T-C	771	992	TpA
	BE9	T-G	298	835	TpC
	CF6	T-C	535	914	TpG
	DC4	T-C	902	1036	TpG
	DC5	T-C	634	947	TpG
	DC5	T-C	648	951	TpG
	DC12	T-C	762	989	TpG
	BF11	T-C	189	798	TpT
	BG5	T-C	85	764	TpT
	BG5	T-C	261	822	TpT
	DA9	T-G	924	1043	TpT
	DD7	T-C	636	947	TpT
3116N	DB7	A-G	653	953	ApA
	DF1	A-G	205	804	ApA
	DG5	A-G	844	1017	ApA
	DH10	A-G	931	1046	ApA
	DC9	A-G	652	953	ApA
	DE8	A-G	662	956	ApC
	DB8	A-G	667	958	ApG
	DE8	A-G	697	968	ApG
	DG2	A-G	697	968	ApG
	DH6	A-G	721	976	ApG
	DB2	A-G	807	1004	ApG
	DD5	A-G	852	1019	ApG
	DA1	A-G	860	1022	ApT
	DF8	A-G	817	1008	ApT
	DC9	A-G	893	1033	ApT
	DG9	A-G	893	1033	ApT
	DD1	C-T	906	1037	CpT
	CA11	G-A	376	861	GpG
	DD8	G-A	850	1019	GpG
	DA10	G-A	782	996	GpT
	DA2	T-C	857	1021	TpA
	DG1	T-C	751	986	TpA
	DH1	T-C	699	968	TpA
	DA8	T-C	695	967	TpC
	DE3	T-C	831	1012	TpC
	DH11	T-C	695	967	TpC
	DB2	T-C	846	1017	TpG
	DE5	T-C	775	994	TpT
3116T	BB8	A-G	109	772	ApC
	BB8	C-T	242	816	CpT
	BD12	G-A	308	838	GpA
	BF4	G-A	31	746	GpG
	BF7	T-C	284	830	TpG
	BD9	T-C	261	822	TpT

Sample ID	Position ID	GPD1 Change	nt	Codon	Target Sequence
3116T	BF7	T-G	260	822	TpT
Continued	CB9	A-G	552	919	ApA
	CF6	A-G	359	855	ApA
	CG4	A-G	500	902	ApA
	DA7	A-G	734	980	ApA
	DH9	A-G	734	980	ApA
	DD5	A-G	854	1020	ApA
	DB5	A-G	927	1044	ApC
	DG3	A-G	864	1023	ApC
	DD11	A-G	735	980	ApC
	CB6	A-G	355	854	ApG
	CG6	A-G	423	876	ApG
	DA9	A-G	929	1045	ApG
	DH6	A-G	899	1035	ApG
	DA7	A-G	680	962	ApT
	BH9	C-T	160	789	CpA
	DC8	C-T	632	946	CpA
	DE5	C-T	832	1013	CpA
	DH9	C-T	928	1045	CpA
	CD6	C-T	450	885	CpC
	CH7	C-T	449	885	CpC
	CG4	C-T	361	856	CpG
	BH10	C-T	99	768	CpT
	CD9	G-A	547	918	GpA
	CG6	G-A	472	893	GpA
	DB11	G-A	763	990	GpA
	DF6	G-A	930	1045	GpA
	DG6	G-A	930	1045	GpA
	DC9	G-A	892	1033	GpA
	CB10	G-A	556	921	GpC
	DD4	G-A	724	977	GpG
	DB10	G-A	704	970	GpT
	DE12	G-A	701	969	GpT
	DG8	G-A	704	970	GpT
	DA2	T-C	806	1004	TpA
	DB12	T-C	861	1022	TpA
	DC12	T-C	726	977	TpC
	CG2	T-C	533	913	TpG
	BH10	T-C	271	826	TpT
	CH9	T-C	427	878	TpT
	DG10	T-C	636	947	TpT
	DD11	T-C	783	996	TpT
1974N	BA3	A-G	253	820	ApC
	BA5	A-G	78	761	ApG
	BA5	A-G	307	838	ApG
	BB4	A-G	128	778	ApT
	BE1	A-G	102	769	ApT
	BB6	C-T	246	817	CpA
	BG5	C-T	151	786	CpG
	BG7	C-T	297	834	CpT
	BH3	T-C	255	820	TpG
	CC2	T-C	444	883	TpG
	BB6	T-C	194	800	TpT
1974T	BD4	A-G	66	757	ApA
	DE10	A-G	619	942	ApA
	DG3	A-G	719	975	ApA
	DH11	A-G	719	975	ApA
	BB8	A-G	75	760	ApG
	BC3	A-G	75	760	ApG
	BC10	A-G	75	760	ApG
	BD1	A-G	75	760	ApG
	BD4	A-G	81	762	ApG
	BG12	A-G	81	762	ApG
	DE2	A-G	677	961	ApT
	DF8	A-G	792	999	ApT
	BC4	C-T	47	751	CpT
	BD6	C-T	47	751	CpT
	BA7	G-A	214	807	GpA
	DH4	G-A	694	967	GpT
	BB4	T-G	181	796	TpA
	BH11	T-C	27	744	TpC
	BH5	T-C	30	745	TpG
533N	A1	A-G	417	874	ApA
	B8	A-G	169	792	ApA
	B3	A-G	140	782	ApG
	A7	A-G	581	929	ApT
	B5	C-T	430	879	CpA
	A7	C-T	289	832	CpC

Sample ID	Position ID	GPD1 Change	nt	Codon	Target Sequence
533N	G9	A-G	359	855	ApA
Continued	G10	A-G	169	792	ApA
	H6	A-G	68	758	ApA
	D1	A-G	280	829	ApC
	D2	A-G	230	812	ApC
	G9	A-G	202	803	ApC
	C8	A-G	69	758	ApG
	D1	A-G	405	870	ApG
	D3	A-G	339	848	ApG
	A9	A-G	508	905	ApG
	D11	A-G	89	765	ApT
	C4	A-G	669	958	ApT
	C7	C-T	168	791	CpA
	H8	C-T	363	856	CpG
	H1	C-T	805	1004	CpT
	G3	G-A	162	789	GpA
	C1	G-A	741	982	GpA
	D9	G-A	799	1002	GpC
	A1	G-A	224	810	GpG
	G5	G-A	283	830	GpT
	G9	T-C	639	948	TpA
	C1	T-C	50	752	TpT
533T	B4	A-G	552	919	ApA
	C10	A-G	66	757	ApA
	C2	A-G	230	812	ApC
	C2	A-G	631	946	ApC
	D7	A-G	864	1023	ApC
	B3	A-G	81	762	ApG
	B7	A-G	95	767	ApG
	D1	A-G	95	767	ApG
	F10	A-G	896	1034	ApG
	D4	A-G	554	920	ApT
	B2	C-T	43	750	CpC
	D7	C-T	738	981	CpT
	E3	G-A	124	777	GpA
	H9	G-A	164	790	GpA
	G5	C-T	802	1003	GpA
	C2	G-A	425	877	GpT
	A7	T-C	699	968	TpA
	E5	T-C	387	864	TpC
	A10	T-C	748	985	TpC
	G2	T-C	889	1032	TpC
2494N	A7	A-G	64	757	ApA
	E2	A-G	312	839	ApA
	H8	A-G	393	866	ApA
	H8	A-G	393	866	ApA
	ME10	A-G	758	988	ApA
	A7	A-G	864	1023	ApC
	MB10	A-G	523	910	ApC
	MA10	A-G	548	918	ApG
	F9	A-G	89	765	ApT
	MF10	A-G	341	849	ApT
	B6	C-T	35	747	CpA
	B9	C-T	201	802	CpA
	A9	G-A	422	876	GpA
	MH9	G-A	837	1014	GpG
	A7	G-A	425	877	GpT
	A9	G-A	716	974	GpT
	B6	T-C	135	780	TpC
	C10	T-C	493	900	TpC
	A10	T-C	748	985	TpC
	B6	T-C	573	926	TpG
	B9	T-C	243	816	TpG
	E10	T-C	30	745	TpG
	F12	T-C	443	883	TpG
	F12	T-C	444	883	TpG
	D5	T-C	316	841	TpT
	E9	T-C	411	872	TpT
2494T	C8	A-G	624	943	ApA
	C2	A-G	380	862	ApC
	D7	A-G	531	912	ApC
	B7	A-G	471	892	ApG
	C2	A-G	161	789	ApG
	B5	A-G	119	775	ApT
	B7	A-G	341	849	ApT
	C12	A-G	119	775	ApT
	C1	C-T	395	867	CpA
	B10	C-T	482	896	CpC

Sample ID	Position ID	GPD1 Change	nt	Codon	Target Sequence
2494T	B10	A-G	733	980	ApA
Continued	D8	A-G	733	980	ApA
	F10	A-G	733	980	ApA
	G4	A-G	380	862	ApC
	F2	A-G	346	851	ApG
	G4	A-G	161	789	ApG
	H3	A-G	161	789	ApG
	B7	A-G	767	991	ApG
	B1	A-G	833	1013	ApT
	C8	A-G	761	989	ApT
	D1	A-G	640	949	ApT
	C4	C-T	290	832	CpC
	D8	C-T	482	896	CpC
	H4	C-T	290	832	CpC
	A2	C-T	880	1029	CpT
	H9	G-A	377	861	GpC
	F11	T-C	34	747	TpC
	B1	T-C	726	977	TpC
	E8	T-C	213	806	TpG
	H4	T-C	714	973	TpG
	C4	T-C	714	973	TpG
	D6	T-C	756	987	TpG
2486N	C8	A-G	118	775	ApA
	MH1	A-G	68	758	ApA
	A2	A-G	125	777	ApC
	C3	A-G	125	777	ApC
	E10	A-G	125	777	ApC
	G4	A-G	607	938	ApC
	G8	A-G	607	938	ApC
	H6	A-G	607	938	ApC
	D10	A-G	607	938	ApC
	G12	A-G	607	938	ApC
	A1	A-G	405	870	ApG
	A11	A-G	405	870	ApG
	B8	A-G	405	870	ApG
	B11	A-G	81	762	ApG
	D4	A-G	161	789	ApG
	D10	A-G	339	848	ApG
	E11	A-G	405	870	ApG
	G4	A-G	339	848	ApG
	G8	A-G	339	848	ApG
	H6	A-G	339	848	ApG
	C12	A-G	405	870	ApG
	C10	C-T	315	840	CpT
	E3	C-T	315	840	CpT
	E4	C-T	373	860	CpT
	H8	C-T	315	840	CpT
	B12	G-A	577	928	GpC
	A5	T-C	789	998	TpC
	C10	T-C	426	877	TpT
	E3	T-C	426	877	TpT
	H8	T-C	426	877	TpT
	C10	T-C	775	994	TpT
2486T	E7	A-G	759	988	ApA
	E3	A-G	116	774	ApC
	H6	A-G	541	916	ApC
	A8	A-G	462	889	ApG
	H3	A-G	341	849	ApT
	B3	A-G	484	897	ApT
	G3	C-T	385	864	CpA
	D7	G-A	106	771	GpA
	E7	G-A	285	830	GpT
	F2	T-C	103	770	TpC
	A8	T-G	539	915	TpG
	A8	T-C	348	851	TpT
2491N	A7	A-G	734	980	ApA
	A3	A-G	660	955	ApT
	C10	A-G	587	931	ApT
	C11	C-T	512	906	CpT
	A2	G-A	526	911	GpA
	B6	G-A	642	949	GpC
	C10	T-C	876	1027	TpA
	D10	T-C	754	987	TpC
	A4	T-C	546	917	TpG
	C3	T-C	576	927	TpG
	A4	T-C	330	845	TpT
	A9	T-C	672	959	TpT
	D3	T-C	330	845	TpT

Sample ID	Position ID	GPD1 Change	nt	Codon	Target Sequence
2491N	MC8	A-C	791	999	ApA
Continued	F2	A-G	541	916	ApC
	F6	A-G	864	1023	ApC
	F2	A-G	431	879	ApG
	F6	A-G	665	957	ApG
	ME7	A-G	565	924	ApG
	G7	A-G	660	955	ApT
	G8	A-G	554	920	ApT
	H12	A-G	660	955	ApT
	MA8	C-T	80	762	CpA
	F6	G-A	787	998	GpG
	E7	T-C	295	834	TpC
	E10	T-C	754	987	TpC
	E6	T-C	576	927	TpG
	G12	T-C	546	917	TpG
	H8	T-C	546	917	TpG
	D4	T-C	330	845	TpT
	E6	T-C	330	845	TpT
	F4	T-C	672	959	TpT
	G12	T-C	330	845	TpT
	H8	T-C	330	845	TpT
2491T	A8	A-G	398	868	ApA
	D6	A-G	625	944	ApA
	D10	A-G	65	757	ApA
	A8	A-G	759	988	ApA
	H6	A-G	735	980	ApC
	A8	A-G	353	853	ApG
	A9	A-G	568	925	ApG
	A10	A-G	292	833	ApG
	B5	A-G	611	939	ApG
	D9	A-G	223	810	ApG
	E8	A-G	309	838	ApG
	B5	A-G	545	917	ApT
	H10	A-G	932	1046	ApT
	H1	C-T	644	950	CpT
	G1	G-A	930	1045	GpA
	A11	G-A	370	859	GpG
	G6	G-A	536	914	GpT
	H4	T-C	279	828	TpA
	A2	T-C	325	844	TpC
	B9	T-C	695	967	TpC
	H8	T-C	135	780	TpC
	H6	T-C	605	937	TpC
	B12	T-C	41	749	TpG
	B12	T-C	634	947	TpG
	C9	T-C	537	914	TpG
	A7	T-C	756	987	TpG
	A10	A-G	840	1015	TpG
	B3	T-C	825	1010	TpG
	D1	T-C	894	1033	TpG
	D10	T-C	634	947	TpG
	F1	T-C	634	947	TpG
	F9	T-C	555	920	TpG
	G3	T-C	825	1010	TpG
	H6	A-G	539	915	TpG
	B2	T-C	684	963	TpT
	D7	T-C	348	851	TpT
	B7	T-C	775	994	TpT
	E8	T-C	316	841	TpT
	H3	T-C	645	950	TpT
2490N	A4	A-G	345	850	ApA
	F4	A-G	157	788	ApA
	G3	A-G	65	757	ApA
	G6	A-G	83	763	ApA
	D7	A-G	380	862	ApC
	E3	A-G	277	828	ApC
	D11	A-G	543	916	ApG
	E7	A-G	119	775	ApT
	D3	C-T	296	834	CpC
	D11	C-T	363	856	CpG
	D7	C-T	299	835	CpT
	G7	G-A	124	777	GpA
	B6	G-A	760	989	GpA
	D11	T-C	641	949	TpG
	F4	T-C	714	973	TpG
	F6	T-C	902	1036	TpG
	D6	T-C	196	801	TpT
	G6	T-C	130	779	TpT

Sample ID	Position ID	APC Change	nt	Codon	Target Sequence
2490N Continued	MC6	A-G	552	919	ApA
	MC6	A-G	202	803	ApC
	ME5	A-G	523	910	ApC
	H3	A-G	913	1040	ApG
	H9	A-G	313	840	ApG
	MA6	A-G	346	851	ApG
	MC6	A-G	667	958	ApG
	F10	A-G	320	842	ApT
	F10	A-G	660	955	ApT
	G5	T-C	806	1004	TpA
	MF6	T-C	295	834	TpC
	G6	T-C	765	990	TpG
	MA6	T-C	444	883	TpG
	F5	T-C	901	1036	TpT
2490T	A8	A-G	62	756	ApA
	B9	A-G	816	1007	ApA
	G5	A-G	631	946	ApC
	A4	A-G	905	1037	ApC
	A8	A-G	794	1000	ApC
	A11	A-G	503	903	ApC
	C9	A-G	548	918	ApG
	D1	A-G	105	770	ApG
	E2	A-G	508	905	ApG
	F6	A-G	81	762	ApG
	D3	A-G	723	976	ApG
	D4	A-G	913	1040	ApG
	E11	A-G	919	1042	ApG
	B9	A-G	554	920	ApT
	C12	A-G	454	887	ApT
	B10	A-G	764	990	ApT
	D3	C-T	532	913	CpA
	A5	G-A	826	1011	GpC
	B8	G-A	915	1040	GpC
	B1	T-C	101	769	TpA
	A1	T-C	325	844	TpC
	B6	T-C	372	859	TpC
	B6	T-G	486	897	TpC
	D8	T-C	325	844	TpC
	A9	T-C	739	982	TpC
	B10	T-C	333	846	TpG
	B9	T-C	825	1010	TpG
	D3	T-C	427	878	TpT
	E4	T-C	40	749	TpT
	B11	T-C	924	1043	TpT
	C11	T-C	753	986	TpT

Appendix C

De novo Mutations

Gene	Exon	Mutation position	Codon	WT codon	Mutant codon	Event	Type	CpG	WT AA	Mutant AA	Ref
APC	4	529	170	GAT	del4b	stop	Fr.		Asp	Fr.	MD
APC	4	529	170	GAT	del4b	stop	Fr.		Asp	Fr.	MD
APC	4	541	174	TTA	ins2b	stop	Fr.		Leu	Fr.	MD
APC	4	550	177	AAT	del1b	stop	Fr.		Asn	Fr.	MD
APC	5	561	181	CAA	TAA	C->T	Ts.	no	Gln	stop	MD
APC	6	666	216	CGA	TGA	C->T	Ts.	yes	Arg	stop	MD
APC	6	666	216	CGA	TGA	C->T	Ts.	yes	Arg	stop	MD
APC	6	714	232	CGA	TGA	C->T	Ts.	yes	Arg	stop	MD
APC	6	749	243	GAG	ins2c	?	?		Glu	?	MD
APC	6	757	232	CGA	TGA	C->T	Ts	Yes	Arg	Stop	81
APC	7	793	244	AGG	del2a	Stop at 250	Fr.		Arg	Fr.	8
APC	8	926	302	CGA	TGA	C->T	Ts.	yes	Arg	stop	MD
APC	10	1374	452	GTT	del2a	stop	Fr.		Val	Fr.	MD
APC	15	1465	482	GAC	ins1b	stop	Fr.		Asp	Fr.	MD
APC	11	1480	473	CAG	del1a	Stop at 497	Fr.		Gln	Fr.	25
APC	11	1555	498	AGA	del1	Stop at 2844	Fr.		Arg	Fr.	81
APC	12	1579	520	TGC	TGA	C->A	Tv.	yes	Cys	stop	MD
APC	13	1680	554	CGA	TGA	C->T	Ts.	yes	Arg	stop	MD
APC	13	1680	554	CGA	TGA	C->T	Ts.	yes	Arg	stop	MD
APC	14	1773	585	CTC	del1a	stop	Fr.		Leu	Fr.	MD
APC	14	1773	591	GCC	del1c	stop	Fr.		Ala	Fr.	MD
APC	14	1895	625	CAG	del4c	stop	Fr.		Gln	Fr.	MD
APC	14	1968	650	GAG	TAG	G->T	Tv.	no	Glu	stop	MD
APC	15	2035	658	GAG	del2c	Stop at 672	Fr.		Glu	Fr.	153
APC	15	2186	722	AGT	del1c	stop	Fr.		Ser	Fr.	MD
APC	15	2277	753	CAT	ins1a	stop	Fr.		His	Fr.	MD
APC	15	2312	764	TTA	Ins1c	stop	Fr.		Leu	Fr.	MD
APC	15	2593	844	TCT	del1a	Stop at 860	Fr.		Ser	Fr.	2
APC	15	2622	868	GAA	TAA	G->T	Tv.	no	Gln	stop	MD
APC	15	2646	876	CGA	TGA	C->T	Ts.	yes	Arg	stop	MD
APC	15	2646	876	CGA	TGA	C->T	Ts.	yes	Arg	stop	MD
APC	15	2689	876	CGA	TGA	C->T	Ts	Yes	Arg	Stop	84
APC	15	2815	932	TCA	TGA	C->G	Tv.	no	Ser	stop	MD
APC	15	2825	935	TAC	del1c	stop	Fr.		Tyr	Fr.	MD
APC	15	2983	974	GGT	ins1c	Stop at 984	Fr.		Gly	Fr.	153
APC	15	3063	1000	TAC	TAA	C->A	Tv	No	Tyr	Stop	153
APC	15	3123	1035	CAG	TAG	C->T	Ts.	yes	Gln	stop	MD
APC	15	3203	1061	AAA	del5c	stop	Fr.		Lys	Fr.	MD

Gene	Exon	Mutation position	Codon	WT codon	Mutant codon	Event	Type	CpG	WT AA	Mutant AA	Ref
APC	15	3203	1061	AAA	del5c	stop	Fr.		Lys	Fr.	MD
APC	15	3203	1061	AAA	del5c	stop	Fr.		Lys	Fr.	MD
APC	15	3209	3209	AGT	del4c	stop	Fr.		Ser	Fr.	MD
APC	15	3220	1067	CAA	del2b	stop	Fr.		Gln	Fr.	MD
APC	15	3226	1055	ATA	del5b	Stop at 1057	Fr.		Ile	Fr.	101
APC	15	3244	1061	AAA	del5a	Stop at 1062	Fr.		Lys	Fr.	7
APC	15	3244	1061	AAA	del5a	Stop at 1062	Fr.		Lys	Fr.	7
APC	15	3244	1061	AAA	del5a	Stop at 1062	Fr.		Lys	Fr.	10
APC	15	3244	1061	AAA	del5a	Stop at 1062	Fr.		Lys	Fr.	13
APC	15	3244	1061	AAA	del5a	Stop at 1062	Fr.		Lys	Fr.	25
APC	15	3244	1061	AAA	del5a	Stop at 1062	Fr.		Lys	Fr.	100
APC	15	3247	1062	CAA	del4a	Stop at 1124	Fr.		Gln	Fr.	81
APC	15	3250	1063	AGT	del4b	Stop at 1124	Fr.		Ser	Fr.	19
APC	15	3259	1066	AGA	del2c	Stop at 1079	Fr.		Arg	Fr.	12
APC	15	3262	1067	CAA	TAA	C->T	Ts	No	Gln	Stop	153
APC	15	3360	1114	CGA	TGA	C->T	Ts.	yes	Arg	stop	MD
APC	15	3394	1125	GTA	ins1b	stop	Fr.		Val	Fr.	MD
APC	15	3394	1125	GTA	ins1b	stop	Fr.		Val	Fr.	MD
APC	15	3571	1170	AAA	ins1	Ins	Fr.		Lys	Fr.	84
APC	15	3595	1192	AAA	del4b	stop	Fr.		Asn	Fr.	MD
APC	15	3598	1193	CAG	del1b	stop	Fr.		Gln	Fr.	MD
APC	15	3838	1259	ATA	del2c	Stop at 1274	Fr.		Ile	Fr.	12
APC	15	3946	1309	GAA	del5b	stop	Fr.		Glu	Fr.	MD
APC	15	3946	1309	GAA	del5b	stop	Fr.		Glu	Fr.	MD
APC	15	3946	1309	GAA	del5b	stop	Fr.		Glu	Fr.	MD
APC	15	3946	1309	GAA	del5b	stop	Fr.		Glu	Fr.	MD
APC	15	3946	1309	GAA	del5b	stop	Fr.		Glu	Fr.	MD
APC	15	3946	1309	GAA	del5b	stop	Fr.		Glu	Fr.	MD
APC	15	3988	1309	GAA	del5	Stop at 2844	Fr.		Glu	Fr.	2
APC	15	3988	1309	GAA	del5	Stop at 2844	Fr.		Glu	Fr.	7
APC	15	3988	1309	GAA	del5	Stop at 2844	Fr.		Glu	Fr.	13
APC	15	3988	1309	GAA	del5	Stop at 2844	Fr.		Glu	Fr.	25

Gene	Exon	Mutation position	Codon	WT codon	Mutant codon	Event	Type	CpG	WT AA	Mutant AA	Ref
APC	15	3988	1309	GAA	del5	Stop at 2844	Fr.		Glu	Fr.	25
APC	15	3988	1309	GAA	del5	Stop at 2844	Fr.		Glu	Fr.	25
APC	15	3988	1309	GAA	del5	Stop at 2844	Fr.		Glu	Fr.	100
APC	15	3988	1309	GAA	del5	Stop at 2844	Fr.		Glu	Fr.	100
APC	15	3988	1309	GAA	del5	Stop at 2844	Fr.		Glu	Fr.	100
APC	15	3988	1309	GAA	del5	Stop at 2844	Fr.		Glu	Fr.	100
APC	15	3988	1309	GAA	del5	Stop at 2844	Fr.		Glu	Fr.	100
APC	15	4033	1324	CCA	del1	Stop at 2844	Fr.		Pro	Fr.	81
APC	15	4152	1378	CAG	TAG	C->T	Ts.	no	Gln	stop	MD
APC	15	4162	1367	CAG	TAG	C->T	Ts	No	Gln	Stop	100
APC	15	4206	1396	TTT	del1a	stop	Fr.		Phe	Fr.	MD
APC	15	4210	1398	AGT	del1a	stop	Fr.		Ser	Fr.	MD
APC	15	4211	1397	GAG	TAG	G->T	Tv.	no	Glu	stop	MD
APC	15	4238	1392	TCA	TAA	C->A	Tv	No	Ser	Stop	153
APC	15	4368	1450	CGA	TGA	C->T	Ts.	yes	Arg	stop	MD
APC	15	4408	1463	AGA	del2a	stop	Fr.		Arg	Fr.	MD
APC	15	4417	1366	GCT	del4b	stop	Fr.		Ala	Fr.	MD
APC	15	4453	1464	GAG	del2a	Stop at 1467	Fr.		Glu	Fr.	100
APC	15	4660	1533	AAT	ins1a	Stop at 1536	Fr.		Asn	Fr.	96
APC	15	5519	1833	AGA	TGA	A->T	Tv.	no	Arg	stop	MD
APC	15	5777	1919	AAT	del16c	stop	Fr.		Asn	Fr.	MD
PAX6	In7	-108			del178						167
PAX6	In7	-108			del178						167
PAX6	In5-2	-19				G->C					167
PAX6	In9	-16			del21						167
PAX6	In8	-6				T->A	Tv				167
PAX6	In11	-2				A->G	Ts				167
PAX6	In11	-2				A->G	Ts				167
PAX6	In5-2	-2				A->G					167
PAX6	In8	-2				C->A	Tv				167
PAX6	In9	-1				G->A	Ts				167
PAX6	In10	1				G->C	Tv				167
PAX6	In6	1				G->A	Ts				167
PAX6	In6	1				G->A	Ts				167
PAX6	In6	1				G->A	Ts				167
PAX6	In6	1				G->C	Tv				167
PAX6	In6	1				G->T	Tv				167

Gene	Exon	Mutation position	Codon	WT codon	Mutant codon	Event	Type	CpG	WT AA	Mutant AA	Ref
PAX6	In6	1				G->T	Tv				167
PAX6	In10	2				T->G	Tv				167
PAX6	In5-1	2				T->C					167
PAX6	In6	2			Ins1	Ins					167
PAX6	In6	2			Ins1	Ins					167
PAX6	Ex5a	20	7	GTC	GAC	T->A	Tv		Val	Asp	167
PAX6	Ex4	363	1	ATG	GTG	A->G	Ts	No	Arg	Val	167
PAX6	Ex5	412	17	AAC	AGC	A->G	Ts	No	Asn	Ser	167
PAX6	Ex5	427	22	CCG	del15	del	in frame		Pro	in frame	167
PAX6	Ex5	447	29	ATT	GTT	A->G	Ts	No	Arg	Val	167
PAX6	Ex5	448	29	ATT	AGT	T->G	Tv	No	Ile	Ser	167
PAX6	Ex5	468	35	AGC	del1		Fr.		Ser	Fr.	167
PAX6	Ex5	471	37	GCC	del9		Fr.		Ala	Fr.	167
PAX6	Ex5	472	37	GCC	ins31	Ins	Fr.		Ala	Fr.	167
PAX6	Ex5	479	39	CGT	del970		Fr.		Arg	Fr.	167
PAX6	Ex5	482	40	TGC	ins41	Ins	Fr.		Cys	Fr.	167
PAX6	Ex5	482	40	GCG	CAG	C->A	Tv	yes	Cys	Gln	167
PAX6	Ex5	482	40	TGC	TGA	C->A	Tv	yes	Arg	Stop	167
PAX6	Ex5	493	44	CGA	CAA	G->A	Ts	yes	Arg	Gln	167
PAX6	Ex5	502	47	CAG	CGG	A->G	Ts	No	Gln	Arg	167
PAX6	Ex6	520	53	GTG	del2	Stop at 54	Fr.		Val	Fr.	167
PAX6	Ex6	526	55	AAA	del7		Fr.		Lys	Fr.	167
PAX6	Ex6	529	56	ATT	ACT	T->C	Ts	No	Arg	Thr	167
PAX6	Ex6	530	56	ATT	del1		Fr.		Tyr	Fr.	167
PAX6	Ex6	545	61	TAC	TAA	C->A	Tv	yes	Arg	Stop	167
PAX6	Ex6	561	67	AGA	TGA	A->T	Tv	No	Pro	Ser	167
PAX6	Ex6	564	68	CCC	TCC	C->T	Ts	No	Gly	Asp	167
PAX6	Ex6	580	73	GGT	GAT	G->A	Ts	No	Ile	Arg	167
PAX6	Ex6	622	87	ATA	AGA	T->G	Tv	No	Arg		167
PAX6	Ex6	626	88	GCC	ins21	Ins	in frame		Arg	in frame	167
PAX6	Ex6	646	95	CCG	ins1	Stop at 116	Fr.		Pro	Fr.	167
PAX6	Ex6	669	103	CGA	TGA	C->T	Ts	yes	Arg	Stop	167
PAX6	Ex6	693	111	CTC	ins1	Ins	Fr.		Leu	Fr.	167
PAX6	Ex6	693	111	CTC	del1	Stop at 124	Fr.		Leu	Fr.	167
PAX6	Ex6	700	113	ACC	ins19	Ins	Fr.		Thr	Fr.	167
PAX6	Ex6	715	118	CCA	ins5	Ins	Fr.		Pro	Fr.	167
PAX6	Ex6	719	119	AGC	AGG	C->G	Tv	yes	Ser	Arg	167
PAX6	Ex7	729	123	ATA	Ins7	Stop at 132	Fr.		Ile	Fr.	167
PAX6	Ex7	765	135	CAA	TAA	C->T	Ts	No	Arg	Stop	167
PAX6	Ex7	765	135	CAA	TAA	C->T	Ts	No	Arg	Stop	167
PAX6	Ex7	790	143	TAT	del4	Stop	Fr.		Tyr	Fr.	167

Gene	Exon	Mutation position	Codon	WT codon	Mutant codon	Event	Type	CpG	WT AA	Mutant AA	Ref
PAX6	Ex7	812	150	AAC	del1	Stop at 186	Fr.		Asn	Fr.	167
PAX6	Ex7	816	152	CAG	TAG	C->T	Ts	No	Arg	Stop	167
PAX6	Ex7	829	156	TGG	del1	Stop at 206	Fr.		Trp	Fr.	167
PAX6	Ex7	862	167	TCG	TGA	CG->GA		yes	Arg	Stop	167
PAX6	Ex8	901	180	CAG	del1	Stop	Fr.		Glu	Fr.	167
PAX6	Ex8	903	181	GAA	TAA	G->T	Tv	No	Arg	Stop	167
PAX6	Ex8	957	199	GAG	TAG	G->T	Tv	No	Arg	Stop	167
PAX6	Ex8	969	203	CGA	TGA	C->T	Ts	yes	Arg	Stop	167
PAX6	Ex8	969	203	CGA	TGA	C->T	Ts	yes	Arg	Stop	167
PAX6	Ex8	969	203	CGA	TGA	C->T	Ts	yes	Arg	Stop	167
PAX6	Ex8	969	203	CGA	TGA	C->T	Ts	yes	Arg	Stop	167
PAX6	Ex8	975	205	CAG	TAG	C->T	Ts	No	Arg	Stop	167
PAX6	Ex9	1051	230	GAG	del1	Stop at 244	Fr.		Glu	Fr.	167
PAX6	Ex9	1071	237	GTG	del2	Stop at 251	Fr.		Val	Fr.	167
PAX6	Ex9	1072	237	GTG	del2	Stop	Fr.		Val	Fr.	167
PAX6	Ex9	1080	240	CGA	TGA	C->T	Ts	yes	Arg	Stop	167
PAX6	Ex9	1080	240	CGA	TGA	C->T	Ts	yes	Arg	Stop	167
PAX6	Ex9	1080	240	CGA	TGA	C->T	Ts	yes	Arg	Stop	167
PAX6	Ex9	1080	240	CGA	TGA	C->T	Ts	yes	Arg	Stop	167
PAX6	Ex9	1080	240	CGA	TGA	C->T	Ts	yes	Arg	Stop	167
PAX6	Ex9	1080	240	CGA	TGA	C->T	Ts	yes	Arg	Stop	167
PAX6	Ex9	1080	240	CGA	TGA	C->T	Ts	yes	Arg	Stop	167
PAX6	Ex9	1103	247	ATA	del11	Stop	Fr.		Ile	Fr.	167
PAX6	Ex9	1127	255	CAG		G->T	Tv	No	Gln	His	167
PAX6	Ex10	1135	258	TTT	TCT	T->C	Ts	No	Phe	Ser	167
PAX6	Ex10	1143	261	CGA	TGA	C->T	Ts	yes	Arg	Stop	167
PAX6	Ex10	1143	261	CGA	TGA	C->T	Ts	yes	Arg	Stop	167
PAX6	Ex10	1143	261	CGA	TGA	C->T	Ts	yes	Arg	Stop	167
PAX6	Ex10	1143	261	CGA	TGA	C->T	Ts	yes	Arg	Stop	167
PAX6	Ex10	1161	267	AGA	TGA	A->T	Tv	yes	Arg	Stop	167
PAX6	Ex10	1180	273	AAT	ins1	Stop	Fr.		Asn	Fr.	167
PAX6	Ex10	1201	280	AAC	del1	Stop	Fr.		Asn	Fr.	167
PAX6	Ex10	1210	283	AGT	ins2	Stop at 317	Fr.		Ser	Fr.	167
PAX6	Ex10	1237	292	AGC	ATC	G->T	Tv	yes	Ser	Ile	167
PAX6	Ex10	1251	297	CAA	TAA	C->T	Ts	No	Arg	Stop	167
PAX6	Ex10	1252	297	CAA	del1	Stop	Fr.		Gln	Fr.	167
PAX6	Ex10	1263	301	CAA	TAA	C->T	Ts	No	Arg	Stop	167
PAX6	Ex11	1311	317	CGA	TGA	C->T	Ts	yes	Arg	Stop	167
PAX6	Ex11	1311	317	CGA	TGA	C->T	Ts	yes	Arg	Stop	167
PAX6	Ex11	1311	317	CGA	TGA	C->T	Ts	yes	Arg	Stop	167
PAX6	Ex11	1311	317	CGA	TGA	C->T	Ts	yes	Arg	Stop	167
PAX6	Ex11	1311	317	CGA	TGA	C->T	Ts	yes	Arg	Stop	167

Gene	Exon	Mutation position	Codon	WT codon	Mutant codon	Event	Type	CpG	WT AA	Mutant AA	Ref
PAX6	Ex12	1426	355	TCC	ins5	Stop	Fr.		Ser	Fr.	167
PAX6	Ex12	1449	363	TCG	CCG	T->C	Ts	No	Ser	Pro	167
PAX6	Ex12	1495	378	CAG	CGG	A->G	Ts	No	Gln	Arg	167
PAX6	Ex12	1503	381	ATG	GTG	A->G	Ts	No	Arg	Val	167
PAX6	Ex12	1513	384	CAG	del4	Stop	Fr.		Gln	Fr.	167
PAX6	Ex12	1533	391	ACC	GCC	A->G	Ts	No	Thr	Ala	167
PAX6	Ex13	1601	413	GAT	del1	Stop	Fr.		Asp	Fr.	167
PAX6	Ex13	1627	422	CAG	CGG	A->G	Ts	No	Gln	Arg	167
PAX6	Ex13	1630	423	TAA	TTA	A->T	Tv	No	Ter	Leu	167
PAX6	Ex13	1630	423	TAA	TTA	A->T	Tv	No	Ter	Leu	167
PAX6	Ex13	1630	423	TAA	TTA	A->T	Tv	No	Ter	Leu	167
PAX6	Ex13	1630	423	TAA	TTA	A->T	Tv	No	Ter	Leu	167
PAX6	Ex13	1630	423	TAA	TTA	A->T	Tv	No	Ter	Leu	167
PAX6	Ex13	1630	423	TAA	TTA	A->T	Tv	No	Ter	Leu	167
HBB	1	125	25	GGT	GTT	G->T	Tv	No	Gly	Val	168
HBB	1	134	28	GCC	GAC	C->A	Tv	No	Ala	Asp	168
HBB	1	134	28	GCC	GAC	C->A	Tv	No	Ala	Asp	168
HBB	1	137	29	CTG	CAG	T->A	Tv	No	Leu	Gln	168
HBB	1	137	29	CTG	CCG	T->C	Ts	No	Leu	Pro	168
HBB	1	137	29	CTG	CCG	T->C	Ts	No	Leu	Pro	168
HBB	1	137	29	CTG	CCG	T->C	Ts	No	leu	Pro	168
HBB	2	149	33	CTG	CGG	T->G	Tv	No	Leu	Arg	168
HBB	2	149	33	CTG	CCG	T->C	Ts	No	Leu	Pro	168
HBB	2	149	33	CTG	CCG	T->C	Ts	No	Leu	Pro	168
HBB	2	178	43	TTT	CTT	T->C	Ts	No	Phe	Leu	168
HBB	2	179	43	TTT	TCT	T->C	Ts	No	Phe	Ser	168
HBB	2	179	43	TTT	TCT	T->C	Ts	No	Phe	Ser	168
HBB	2	179	43	TTT	TCT	T->C	Ts	No	Phe	Ser	168
HBB	2	179	43	TTT	TCT	T->C	Ts	No	Phe	Ser	168
HBB	2	241	64	CAT	TAT	C->T	Ts	No	His	Tyr	168
HBB	2	241	64	CAT	TAT	C->T	Ts	No	His	Tyr	168
HBB	2	241	64	CAT	TAT	C->T	Ts	No	His	Tyr	168
HBB	2	241	64	CAT	TAT	C->T	Ts	No	His	Tyr	168
HBB	2	241	64	CAT	TAT	C->T	Ts	No	His	Tyr	168
HBB	2	241	64	CAT	TAT	C->T	Ts	No	His	Tyr	168
HBB	2	241	64	CAT	TAT	C->T	Ts	No	His	Tyr	168
HBB	2	242	64	CAT	CCT	A->C	Tv	No	His	Pro	168
HBB	2	247	99	GTG	GGG	T->G	Tv	No	Val	Gly	168
HBB	2	250	67	AAA	GAA	A->G	Ts	No	Lys	Glu	168
HBB	2	254	68	GTG	GAG	T->A	Tv	No	Val	Asp	168
HBB	2	254	68	GTG	CAG	T->A	Tv	No	Val	Asp	168
HBB	2	254	68	GTG	GAG	T->A	Tv	No	Val	Glu	168
HBB	2	257	69	CTC	CCC	T->C	Ts	No	Leu	Pro	168
HBB	2	266	72	TTT	TCT	T->C	Ts	No	Phe	Ser	168
HBB	2	296	82	CTC	CGC	T->G	Tv	No	Leu	Arg	168
HBB	2	308	86	TTT	TCT	T->C	Ts	No	Phe	Ser	168
HBB	2	317	89	CTG	CCG	T->C	Ts	No	Leu	Pro	168

Gene	Exon	Mutation position	Codon	WT codon	Mutant codon	Event	Type	CpG	WT AA	Mutant AA	Ref
HBB	2	317	89	CTG	CCG	T->C	Ts	No	Leu	Pro	168
HBB	2	326	92	CTG	CCG	T->C	Ts	No	Leu	Pro	168
HBB	2	328	93	CAC	TAC	C->T	Ts	No	His	Tyr	168
HBB	2	328	93	CAC	TAC	C->T	Ts	No	His	Tyr	168
HBB	2	330	93	CAC	CAG/A	C->A/G	Tv	No	His	Gln	168
HBB	2	346	99	GTG	ATG	G->A	Ts	Yes	Val	Met	168
HBB	2	346	99	GTG	ATG	G->A	Ts	Yes	Val	Met	168
HBB	2	346	99	GTG	ATG	G->A	Ts	Yes	Val	Met	168
HBB	2	346	99	GTG	ATG	G->A	Ts	Yes	Val	Met	168
HBB	2	346	99	GTG	ATG	G->A	Ts	Yes	Val	Met	168
HBB	3	371	107	CTG	CCG	T->C	Ts	No	Leu	Pro	168
HBB	3	371	107	CTG	CCG	T->C	Ts	No	Leu	Pro	168
HBB	3	371	107	CTG	CCG	T->C	Ts	No	Leu	Pro	168
HBB	3	371	107	CTG	CCG	T->C	Ts	No	Leu	Pro	168
HBB	3	397	116	GCC	CCC	G->C	Tv	No	Ala	Pro	168
HBB	3	476	142	CTG	CGG	T->G	Tv	No	Leu	Arg	168
HBB	3	478	143	GCC	CCC	G->C	Tv	No	Ala	Pro	168

References for *de novo* Mutations

References 2-153 taken from *APC* mutation database:

2 Olschwang S, Tiret A, Laurent-Puig P, Muleris M, Parc R and Thomas G (1993) Restriction of ocular fundus lesions to a specific subgroup of APC mutations in adenomatous polyposis coli patients. Cell 75: 959-968.

7 Groden J, Gelbert L, Thliveris A, Nelson L, Robertson M, Joslyn G, Samowitz W, Spirio L, Carlson M, Burt R and et al. (1993) Mutational analysis of patients with adenomatous polyposis: identical inactivating mutations in unrelated individuals. Am J Hum Genet 52: 263-72.

8 Groden, J., Thliveris, A., Samowitz, W., Carlson, M., Gelbert, L., Albertsen, H., Joslyn, G., Stevens, J., Spirio, L., Robertson, M., and et al. (1991). Identification and characterization of the familial adenomatous polyposis coli gene. Cell 66, 589-600.

10 Varesco L, Gismondi V, James R, Robertson M, Grammatico P, Groden J, Casarino L, De Benedetti L, Bafico A, Bertario L and et al. (1993) Identification of APC gene mutations in Italian adenomatous polyposis coli patients by PCR-SSCP analysis. Am J Hum Genet 52: 280-5.

12 Bapat B, Berk T, Mitri A, Cohen Z, Gallinger S and Stern H (1994) Characterization of two novel adenomatous polyposis coli (APC) gene mutations in patients with familial adenomatous polyposis (FAP). Human Mutation 4: 253-256.

13 Cama A, Palmirotta R, Curia MC, Esposito DL, Ranieri A, Ficari F, Valanzano R, Battista P, Modesti A, Tonelli F and et al. (1995) Multiplex PCR analysis and genotype-phenotype correlations of frequent APC mutations. Hum Mutat 5: 144-52.

19 Ichii S, Nagase H, Mori T, Baba S and Nakamura Y (1993) A novel mutation of the APC (adenomatous polyposis coli) gene in a familial adenomatous polyposis (FAP) patient and presymptomatic diagnosis using PCR. Hum Mol Genet 2: 597.

25 Mandl M, Paffenholz R, Friedl W, Caspari R, Sengteller M and Propping P (1994) Frequency of common and novel inactivating APC mutations in 202 families with familial adenomatous polyposis. Hum Mol Genet 3: 181-184.

81 Dobbie Z, Spycher M, Mary JL, Haner M, Guldenschuh I, Hurliman R, Amman R, Roth J, Muller H and Scott RJ (1996) Correlation between the development of extracolonic manifestations in FAP patients and mutations beyond codon 1403 in the APC gene. J Med Genet 33: 274-80.

84 Nordling M, Engwall Y, Wahlstrom J, Wiklund L, Eriksson MA, Gustavsson B, Fasth S, Larsson PA and Martinsson T (1997) Novel mutations in the APC gene and clinical features in Swedish patients with polyposis coli. Anticancer Res 17: 4275-4280.

96 Resta N, Stella A, Susca F, Montera M, Gentile M, Cariola F, Prete F, Tenconi R, Tibiletti MG, Logrieco G, Mattina T, Andriulli G, Caruso ML, Fiorente P, Russo S, Caputi-Jambrenghi O, Mareni C and Guanti G (2001) Nine novel APC mutations in Italian FAP patients. Hum Mutat 17: 434-5.

100 Gayther SA, Wells D, SenGupta SB, Chapman P, Neale K, Tsioupra K and Delhanty JD (1994) Regionally clustered APC mutations are associated with a severe phenotype and occur at a high frequency in new mutation cases of adenomatous polyposis coli. Hum Mol Genet 3: 53-6.

101 Mandl M, Kadmon M, Sengteller M, Caspari R, Propping P and Friedl W (1994) A somatic mutation in the adenomatous polyposis coli (APC) gene in peripheral blood cells--implications for predictive diagnosis. Hum Mol Genet 3: 1009-11.

153 Giarola M, Stagi L, Presciuttini S, Mondini P, Radice MT, Sala P, Pierotti MA, Bertario L and Radice P (1999) Screening for mutations of the APC gene in 66 Italian familial adenomatous polyposis patients: evidence for phenotypic differences in cases with and without identified mutation. Hum Mutat 13: 116-23.

167 http://pax6.hgu.mrc.ac.uk/

168 Stamatoyannopoulos,G. and Nute,P.E. (1982). De novo mutations producing unstable Hbs or Hbs M. II. Direct estimates of minimum nucleotide mutation rates in man. Hum. Genet. *60*, 181-188.

MD Patients from Scottish polyposis register.